A.D. MYŠKIS

ADVANCED MATHEMATICS *FOR* ENGINEERS

special courses

MIR
PUBLISHERS

ADVANCED MATHEMATICS FOR ENGINEERS

MIR PUBLISHERS

А. Д. МЫШКИС

МАТЕМАТИКА ДЛЯ ВТУЗОВ

СПЕЦИАЛЬНЫЕ КУРСЫ

ИЗДАТЕЛЬСТВО «НАУКА» МОСКВА

A. D. MYŠKIS

ADVANCED MATHEMATICS FOR ENGINEERS

Special Courses

TRANSLATED FROM THE RUSSIAN

BY

V. M. VOLOSOV, D. Sc.

AND

I. G. VOLOSOVA

MIR PUBLISHERS MOSCOW

First published 1975

Revised from the 1971 Russian edition

На английском языке

© English translation, Mir Publishers, 1975

Preface

The present book is a collection of special courses in higher mathematics intended for advanced engineering students, physicists and applied scientists working in academic and industrial research. It stems from lectures given by the author for a number of years and can be used both for studying under the guidance of a teacher and for self-instruction.

The reader is supposed to be familiar with fundamentals of higher mathematics (for this purpose we can recommend [107]). In the presentation of the material primary emphasis has been placed on the practical significance of mathematical notions and their application. This is a course of applied mathematics and not a theorem-proving book, and therefore intuitive methods take precedence over logical ones.

In every division of mathematics treated in this course we confine ourselves to a necessary minimum of basic ideas, theoretical facts and applications. For more detail the reader is referred to supplementary books. The asterisk in the table of contents indicates the sections which can be omitted in the first acquaintance with the subject-matter.

The sections entering into each chapter are numbered in succession beginning with the first number. In references inside each chapter or section the number of the chapter or section is omitted. For instance, the expression "formula (2)" placed in § 3 of Chapter IV means "formula (2) in § 3 of Chapter IV", the expression "formula (1.2)" in the text of Chapter IV means "formula (2) in § 1 of Chapter IV" and "formula (III.4.2.)" means "formula (2)

in § 4 of Chapter III", "Sec. IV.2.3" means "Section 3 of § 2 in Chapter IV", etc.

While preparing the book for publication the author has received valuable advice and criticism from R. S. Guter, N. D. Kopachevsky, M. A. Krasnoselsky, A. D. Tyuptsov, G. L. Lunts, A. G. Mladov and others. To all of them the author expresses his warmest gratitude. *)

A. D. Myškis

*) The present translation incorporates suggestions made by the author.—*Tr.*

Contents

Preface 5

CHAPTER I. SCALAR AND VECTOR FIELDS 15

§ 1. Hamiltonian Operator 15

1. Operations with Symbol ∇ 16
2. Rules of Application 17
3. Integral Formulas 19
4. Repeated Operations with ∇ 20
5. Discontinuous Fields 21

§ 2. Special Types of Field 23

1. Potential Fields 23
2. Irrotational Field in Multiply Connected Domain 24
3. Solenoidal Fields 26
4. Examples 29
5. Newtonian Potential 31
6. Constructing Vector Field with Given Rotation and Divergence . 33

CHAPTER II. THE THEORY OF ANALYTIC FUNCTIONS 36

§ 1. Differentiation and Mappings 36

1. Derivative 36
2. Cauchy-Riemann Equations 37
3. Conjugate Harmonic Functions 38
4. Geometric Interpretation of Derivative 39
5. Conformal Mappings 41
6. Linear Mappings 42
7. Extended Complex Plane 43
8. Fractional Linear Mapping 44
9. The Mapping $w=z^k$ 48
10. Multiple-Valued Functions and Branch Points 50
11*[1]. The Mapping $w=\frac{c}{2}\left(z+\frac{1}{z}\right)$ 54
12. Exponential Mapping and Some Mappings Related to It 57
13*. Riemann Surface 59
14. Application to the Theory of Plane Fields 61
15. Examples 64
16*. Boundary-Value Problems and Conformal Mappings 66

[1] The items starred in the table of contents indicate sections that contain optional material which may be omitted in a first reading of the book.

17. General Remarks on Conformal Mappings 71
18*. Application of Small-Parameter Method 74

§ 2. Integration and Power Series 78

1. Integral 78
2. Integral of an Analytic Function 79
3. Laurent Series 80
4. Expanding an Analytic Function into a Laurent Series 82
5. Taylor Series 85
6. Analytic Mappings and Maximum Modulus Principle 88
7. Analytic Continuation 91
8. Some Generalizations 94

§ 3. Singular Points and Zeros 100

1. Isolated Singular Points 100
2. Pole 101
3. Cauchy's Residue Theorem 103
4. Application to Improper Integrals 106
5. Poisson Integral 115
6. Behaviour of a Function at Infinity 118
7. Logarithmic Residues 120
8. Rouché's Theorem 122
9. Zeros Dependent on a Parameter 123
10. Zeros of Polynomials 126
11*. Resultant of Two Polynomial Equations 130
12*. Meromorphic Functions 132
13*. Schwarz-Christoffel Transformation 136
14*. Elliptic Functions 141

§ 4. Asymptotic Expansions 144

1. Introduction 144
2. Properties 146
3. The Fourier-Type Integral 149
4. Integral of Type $I(\nu)=\int_a^b f(x)\,e^{\nu\varphi(x)}\,dx$ 153
5. Saddle-Point Method 156

CHAPTER III. OPERATIONAL CALCULUS 162

§ 1. General Theory 162

1. Laplace Transformation 162
2. Transforms of Simple Functions 164
3. Basic Properties of the Laplace Transformation 166
4. Laplace's Inverse Transformation 170
5. Expanding the Original into a Series 174
6*. Numerical Computation of the Original 177

§ 2. Applications 180

1. Basic Idea 180
2. Ordinary Differential Equations 181
3. Difference and Difference-Differential Equations 185
4. Integral and Integro-Differential Equations 187
5. Partial Differential Equations 188

§ 3. Some Generalizations* 192

1*. Discrete Laplace Transformation 193
2*. Fourier Transformation of Functions of Bounded Order of Growth 195
3*. Some Other Integral Transformations 197
4*. Integral Transformations on a Finite Interval 201

CHAPTER IV. LINEAR ALGEBRA 204

§ 1. Conjugate Mappings 206

1. Direct Sum 206
2. Invariant Subspaces 206
3. Conjugate Mappings 208
4. Decomposition Connected with Conjugate Mappings 209
5. Mapping of a Space into Itself 210
6. Self-Adjoint Mapping 212
7. Extremal Properties of Eigenvalues 213

§ 2. Quadratic Forms 216

1. Introduction 216
2. Law of Inertia for Quadratic Forms 218
3. Jacobi Method and Sylvester Theorem 219
4. Simultaneous Reduction of Two Quadratic Forms to Canonical Form 220

§ 3. Structure of Linear Mapping 222

1. Mapping with a Single Eigenvector 222
2. Mapping with a Single Eigenvalue 225
3. General Case 227
4. Mapping of a Real Space 230
5. Application to Constructing Functions of Matrices 233
6. Another Representation of a Mapping of a Real Space 236
7. Commuting Mappings 237

§ 4. Numerical Methods 238

1. Gauss' Method 238
2. Norm of a Matrix 241
3. Method of Minimizing Discrepancy 243
4. Spectrum of a Symmetric Matrix 244
5. Method of K.G.J. Jacobi 245
6. Iteration Method for Computing Eigenvalues with the Greatest Modulus 246
7. Computing Subsequent Eigenvalues 248
8*. Matrices with Nonnegative Elements 250
9. Method of A. N. Krylov 252
10. Small-Parameter Method 253
11. Method of Continuous Extension 254

§ 5. Linear Programming 256

1. Basic Problem 256
2. Examples 258
3. Geometric Considerations 259
4. Geometric Interpretation of the Basic Problem 262
5. Standard Form of the Basic Problem 264
6. Simplex Method 265
7. Application to Matrix Games 270
8. Some Other Problems 276

CHAPTER V. TENSORS 280

§ 1. Tensor Algebra 281

1. Examples 281
2. Cartesian Tensors. General Definition 284
3. Operations on Tensors 285
4. Tensors of Second Rank 287
5. Application to Mechanics 288
6. General Affine Tensors 291
7. Affine Tensors in a Euclidean Space 294
8*. Pseudo-Euclidean Spaces 295
9. Remark on Dimensions 299

§ 2. Tensor Fields 300

1. Field of a Cartesian Tensor 300
2. Christoffel Symbols 302
3. Covariant Differentiation 305
4. Tensor Field on a Manifold of a Euclidean Space 308
5*. Intrinsic Geometry and Riemannian Spaces 310

CHAPTER VI. CALCULUS OF VARIATIONS 317

§ 1. First Variation and Sufficient Conditions for Extremum 317

1. Examples of Variational Problems 317
2. Functional 320
3. Functional Spaces 321
4. Variation of a Functional 326
5*. More Precise Definition 329
6. Necessary Condition for Extremum 331
7. Euler's Equation 333
8. Examples 336
9. Functionals Involving Derivatives of Higher Orders 338
10. Functionals of Several Functions 339
11. Functionals of Functions of Several Variables 341
12. Isoperimetric Problems 344
13. Conditional Extremum with Finite and Differential Constraints . 348
14*. Problems Reducible to Lagrange's Problem 351
15. Extremals with Moving Boundaries in the Plane 353
16*. Transversality Conditions 354
17*. Extremals with Moving Boundaries in Space 357
18*. Transversality Conditions for Functions of Several Variables . . 361
19*. Unilateral Constraints 363
20. Extremals with Corners 365

§ 2. Second Variation and Sufficient Conditions for Extremum 369

1. Variations of Higher Order 369
2. Expressing Sufficient Conditions for Extremum in Terms of Second Variation 371
3. Legendre's Necessary Conditions 372
4. Quadratic Functional 374
5. Jacobi Conditions 377
6. Geodesics 381
7. Conditions for Strong Extremum 385
8. Variational Theory of Eigenvalues 387
9*. Existence of Minimum 393
10. Basic Condition for Minimum 396
11. Dependence of Eigenvalues on the Functional 399

§ 3. Canonical Equations and Variational Principles 401

1. Canonical Form of Euler's Equations 401
2. First Integrals 403
3. Canonical Transformations 405
4*. Contact Transformations 407
5*. Noether Theorem 410
6*. The Case of Functions of Several Variables 413
7. Hamilton-Jacobi Equation 415
8*. Lobachevskian Plane 418
9. Variational Principles 421
10. Hamilton's Principle. The Simplest Case 423
11. Hamilton's Principle for Systems with Finite Number of Degrees of Freedom 426
12*. Hamilton's Principle for Continuous Media. Vibrations of a String 429
13*. Vibrations of a Bar and of a Plate 433
14*. General Scheme for Applying Variational Principles to Physical Fields 437
15. Equations of Motion of an Elastic Medium 442
16. Dissipative Systems 443
17. Principle of Minimum of the Potential Energy 445
18*. Examples 446
19. Barrier Potential 449
20*. Variational Principle for Conformal Mappings 452

§ 4. Direct Methods 454

1. The Ritz Method for a Quadratic Functional 455
2*. Application to Boundary-Value Problems 462
3*. Infinite Sequence of Coordinate Functions 463
4. The Ritz Method for Functionals of Functions of Several Variables 465
5. The Trefftz Method 471
6. The Ritz Method for Computing Eigenvalues 472
7*. The Ritz Method for General Functionals 475
8. Method of Least Squares 479
9*. Method of L. V. Kantorovich 481
10. Euler's Method 484

CHAPTER VII. INTEGRAL EQUATIONS 486

§ 1. Introduction 486

1. Examples 486
2. Basic Classes of Integral Equations 488
3. More on the Hilbert Space 490

§ 2. Fredholm Theory 491

1. Equations with Degenerate Kernels 491
2. General Case 497
3. Applying Infinite Systems of Algebraic Equations 501
4. Applying Numerical Integration 504
5. Equations with Small Kernels 509
6*. Contraction Mapping Principle 511
7. Perturbed Kernels 515
8. Properties of Solutions 517
9. Volterra's Integral Equations of the Second Kind 519
0*. Equations with Polar Singularities 521
11*. Equations with Completely Continuous Operators 523
12*. Equations with Positive Kernels 524

§ 3. Equations with Symmetric Kernels 525

1. Finite-Dimensional Analogue 525
2. Expanding Kernels with Respect to Eigenfunctions 526
3. Corollaries 528
4*. Passing from Nonsymmetric Kernel to Symmetric Kernel 533
5. Extremal Property of Characteristic Values 536
6*. Equations with Self-Adjoint Operators 539

§ 4. Special Classes of Integral Equations 542

1. Volterra's Equation of the First Kind 542
2. Fredholm's Equations of the First Kind with Symmetric Kernel . 545
3*. Improperly Posed Problems 547
4. Fredholm's Equations of the First Kind. General Case 548
5*. Generating Functions 550
6. Volterra's Equations with Difference Kernels 555
7*. Fredholm's One-Dimensional Equations with Difference Kernels . . 557
8*. Fredholm's One-Dimensional Equations with Difference Kernels on the Semi-Axis 564

§ 5. Singular Integral Equations 569

1. Singular Integrals 569
2. Inversion Formulas 572
3. Direct Application of Inversion Formulas 574
4. Reduction to a Boundary-Value Problem. Example 576
5. The Case of an Arbitrary Closed Contour 578
6. The Case of a Nonclosed Contour 583
7*. Reduction to an Infinite System of Algebraic Equations 585

§ 6. Nonlinear Integral Equations 587

1. Reduction to Finite Equations 587
2. Iteration Method 590
3. Small-Parameter Method 592
4*. Application of the Theory of Symmetric Kernels 593
5*. Application of Fixed-Point Theorems 595
6*. Variational Methods 598
7. Equations with Parameters 599
8. Bifurcation of Solutions 601

CHAPTER VIII. ORDINARY DIFFERENTIAL EQUATIONS 606

§ 1. Linear Equations and Systems 606

1. General Properties 606
2. Periodic Systems 611
3*. Hill's Equation 614
4. Parametric Resonance 620
5. Hamiltonian Systems 621
6. Nonhomogeneous Systems 624
7*. Almost Periodic Functions 627
8. Asymptotic Expansions of Solutions for $t \to \infty$ 629
9. More on Asymptotic Behaviour of Solutions 633
10. Oscillation of Solutions of Second-Order Equations 636
11. Systems Involving Parameters 640
12*. Turning Points 644

§ 2. Autonomous Systems 647

1. General Concepts 647
2. Limiting Behaviour of Solutions 648
3. Rest Points in the Plane. Linear Systems 649
4. General Case 654
5. Cycles in the Plane 658
6. Index of a Singular Point of a Vector Field 661
7. Rest Points in Space 664
8. Cycles in Space 667
9*. Structurally Stable Systems 671
10. Discontinuous Systems 672
11. Systems on Manifolds 676
12. Systems with an Integral Invariant 678
13. Ergodicity 680

§ 3. Stability of Solutions 684

1. Introduction 684
2. First-Order Equations 686
3. Lyapunov's Function 688
4. Stability in Linear Approximation 692
5*. Critical Cases 698
6. Special Classes of Mechanical Systems 705
7. Automatic Control Systems 713
8*. "Technical" Stability 718

§ 4. Nonlinear Vibrations 720

1. Introduction 720
2. Free Oscillations of a Conservative Autonomous System with One Degree of Freedom 727
3. Forced Oscillations of a Weakly Nonlinear System. Basic Case . . 734
4. Singular Cases 736
5. Subharmonic Oscillations 741
6. More on Forced Oscillations 743
7. Auto-Oscillations 746
8. Relaxation Oscillations 749
9*. Boundary Layer 752
10. Nonperiodic Oscillations 755
11*. Krylov-Bogolyubov Method 762
12. Discrete-Time Systems 765

Bibliography 770

Name Index 776

Subject Index 780

List of Symbols 791

CHAPTER I

Scalar and Vector Fields

We suppose that the reader is familiar with the basic notions of the theory of scalar and vector fields such as directional derivative, gradient, vector line, flux of a vector, divergence, circulation and rotation (e.g. see [107], Secs. IX.9, XII.1, 2 and 4, XVI.21-27). Here we shall present some more facts of the theory. For more detail the reader is referred to [14] and [64].

§ 1. HAMILTONIAN OPERATOR

We remind the reader that we associate with a scalar field u the vector field of its gradient

$$\text{grad } u = \frac{\partial u}{\partial x}\mathbf{i} + \frac{\partial u}{\partial y}\mathbf{j} + \frac{\partial u}{\partial z}\mathbf{k} \tag{1}$$

where x, y and z are Cartesian coordinates. The gradient can also be given an invariant-definition independent of the particular choice of the coordinate system. This can be done on the basis of the notion of a directional derivative, and therefore here and henceforth we consider the gradient of a vector field and other "derived fields" as connected with the given field in an invariant manner.

With a given vector field $\mathbf{A}$ we also associate the scalar field of its **divergence**

$$\text{div } \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} \tag{2}$$

and the vector field of its **rotation**

$$\text{rot } \mathbf{A} = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right)\mathbf{i} + \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right)\mathbf{j} + \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)\mathbf{k} \tag{3}$$

The divergence and rotation can also be given invariant definitions: the former on the basis of the notion of a *flux* and the latter on the basis of *circulation*. The divergence enters into one of the

basic formulas of vector analysis, namely into the **Ostrogradsky-Gauss** *) formula for the **flux** of a vector (or, more precisely, of a vector field) through a closed surface (σ) which is defined as the integral $\oint\limits_{(\sigma)} \mathbf{A}\cdot\mathbf{d\sigma} = \oint\limits_{(\sigma)} A_n\, d\sigma$:

$$\oint\limits_{(\sigma)} \mathbf{A}\cdot\mathbf{d\sigma} = \int\limits_{(\Omega)} \operatorname{div}\mathbf{A}\, d\Omega \tag{4}$$

Here (Ω) is the domain (region) bounded by the surface (σ) and A_n is the projection of $\mathbf{A}$ on the outer normal to (σ) which goes in the direction from the interior of (Ω) to the surrounding space. The rotation enters into another important formula known as **Stokes' formula** which expresses the **circulation** of a vector field over a closed oriented contour (L); the circulation is defined as the integral $\oint\limits_{(L)} \mathbf{A}\cdot d\mathbf{r} = \oint\limits_{(L)} A_\tau\, dL$ and is expressed in terms of a surface integral by the formula

$$\oint\limits_{(L)} \mathbf{A}\cdot d\mathbf{r} = \int\limits_{(S)} \operatorname{rot}\mathbf{A}\cdot\mathbf{dS}$$

where $\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ is the radius vector, τ is the unit tangent vector to (L) in the direction of traversing the contour, and (S) is an arbitrary surface bounded by the contour (L) ("pulled over (L)") and coherently oriented with (L) (e.g. see [107], Sec. VII.11).

For plane fields these formulas are simplified accordingly.

For a given vector field we also consider its **vector lines** which fill the whole region occupied by the field and whose tangent at each point goes in the direction of the field.

1. Operations with Symbol ∇. To simplify the expressions of operations (1), (2) and (3) R.W. Hamilton (a noted Irish mathematician, 1805-1865) introduced the symbolic vector operator

$$\nabla = \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}$$

read **nabla** (or **del**). The term nabla originates from the Greek νάβλα which is the name of an ancient musical instrument triangular in shape. This symbol designates an operation on a field, i.e. it is an **operator** (for the notion of an operator we can refer the reader to [107], Secs. XIV.26 and XV.20). We say that the **Hamiltonian operator** ∇ is a *vector differential operator* because it possesses properties both of a vector and of the operator of differentiation.

*) Gauss, Karl Friedrich (1777-1855), a great German mathematician. Ostrogradsky, Mikhail Vasilyevich (1801-1862), a celebrated Russian mathematician.—*Tr.*

The application of the Hamiltonian operator to a scalar (or, more precisely, to a scalar field) u and to a vector $\mathbf{A}$ is formally written as symbolic multiplication and is performed according to the following rules:

$$\nabla u = \left(\mathbf{i}\frac{\partial}{\partial x}+\mathbf{j}\frac{\partial}{\partial y}+\mathbf{k}\frac{\partial}{\partial z}\right)u = \mathbf{i}\frac{\partial u}{\partial x}+\mathbf{j}\frac{\partial u}{\partial y}+\mathbf{k}\frac{\partial u}{\partial z} = \operatorname{grad} u \tag{5}$$

$$\nabla\cdot\mathbf{A} = \left(\mathbf{i}\frac{\partial}{\partial x}+\mathbf{j}\frac{\partial}{\partial y}+\mathbf{k}\frac{\partial}{\partial z}\right)\cdot(\mathbf{i}A_x+\mathbf{j}A_y+\mathbf{k}A_z) = $$

$$= \frac{\partial A_x}{\partial x}+\frac{\partial A_y}{\partial y}+\frac{\partial A_z}{\partial z} = \operatorname{div}\mathbf{A} \tag{6}$$

$$\nabla\times\mathbf{A} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ A_x & A_y & A_z \end{vmatrix} = \mathbf{i}\left(\frac{\partial A_z}{\partial y}-\frac{\partial A_y}{\partial z}\right)+\mathbf{j}\left(\frac{\partial A_x}{\partial z}-\frac{\partial A_z}{\partial x}\right)+$$

$$+\mathbf{k}\left(\frac{\partial A_y}{\partial x}-\frac{\partial A_x}{\partial y}\right) = \operatorname{rot}\mathbf{A} \tag{7}$$

The operator ∇ is applied as a differential operator only to the factor directly following it. The result of such an operation is a concrete quantity and is no longer applied as an operator. Therefore, one must avoid writing ambiguous expressions of the type ∇uv; these should be written in the form $(\nabla u)v = (\operatorname{grad} u)v = v\operatorname{grad} u$ or $\nabla(uv) = \operatorname{grad}(uv)$ (which are different things!) depending on the meaning of the expression.

If ∇ is not followed by a vector or scalar factor, then the corresponding expression is understood as an operator. For instance, the expression

$$\mathbf{A}\cdot\nabla = (\mathbf{i}A_x+\mathbf{j}A_y+\mathbf{k}A_z)\cdot\left(\mathbf{i}\frac{\partial}{\partial x}+\mathbf{j}\frac{\partial}{\partial y}+\mathbf{k}\frac{\partial}{\partial z}\right) =$$

$$= A_x\frac{\partial}{\partial x}+A_y\frac{\partial}{\partial y}+A_z\frac{\partial}{\partial z}$$

is a scalar differential operator which can be applied both to vector and scalar fields. Using the well-known formula for the directional derivative we can also write this operator in the form $A\frac{\partial}{\partial l_{\mathbf{A}}}(A=|\mathbf{A}|)$ where $l_{\mathbf{A}}$ is the direction of the vector $\mathbf{A}$. In particular, the rate of change of a scalar or vector field along a trajectory (e.g. see [107], Sec. XII.1) can be expressed by means of the operator relation

$$\frac{d}{dt} = \mathbf{v}\cdot\nabla+\frac{\partial}{\partial t}$$

which is applied to the given field.

2. Rules of Application. When manipulating the symbolic operator ∇ we use the rules of vector algebra and the rules of

differentiation. For example, we have

$$\operatorname{rot}(\mathbf{A}+\mathbf{B}) = \nabla \times (\mathbf{A}+\mathbf{B}) = \nabla \times \mathbf{A} + \nabla \times \mathbf{B} = \operatorname{rot}\mathbf{A} + \operatorname{rot}\mathbf{B}$$
$$\operatorname{grad}(\lambda u) = \nabla(\lambda u) = \lambda \nabla u = \lambda \operatorname{grad} u \quad (\lambda = \text{const}) \tag{8}$$

because multiplication by a scalar and differentiation possess the linearity property. At the same time we cannot regard λ in formula (8) as a point function in space, i. e. as a scalar field, because a variable quantity cannot be taken outside the sign of differentiation. To include the latter case we note that in the ordinary formula

$$(uv)' = u'v + uv' \tag{9}$$

for differentiating a product the first term is obtained if we regard v as a constant when performing differentiation and the second term if u is considered to be constant. Therefore we can rewrite formula (9) in the form

$$(uv)' = (u_c v)' + (uv_c)' = u_c v' + u' v_c = uv' + u'v$$

where the subscript c (which means "constant") indicates that while performing differentiation we assume the corresponding quantity to be constant. (Of course, when this quantity is not under the sign of differentiation the subscript can be omitted.) Thus, using this convention, we can write

$$\operatorname{grad}(uv) = \nabla(uv) = \nabla(u_c v) + \nabla(uv_c) = u\nabla v + v\nabla u = u \operatorname{grad} v + v \operatorname{grad} u$$

Here are further examples of operations with ∇:

$$\operatorname{div}(u\mathbf{A}) = \nabla \cdot (u\mathbf{A}) = \nabla \cdot (u_c \mathbf{A}) + \nabla \cdot (u\mathbf{A}_c) = u(\nabla \cdot \mathbf{A}) + \nabla u \cdot \mathbf{A} =$$
$$= u \operatorname{div}\mathbf{A} + \operatorname{grad} u \cdot \mathbf{A} \tag{10}$$
$$\operatorname{rot}(\mathbf{A} \times \mathbf{B}) = \mathbf{A} \operatorname{div}\mathbf{B} - \mathbf{B} \operatorname{div}\mathbf{A} + (\mathbf{B}\cdot\nabla)\mathbf{A} - (\mathbf{A}\cdot\nabla)\mathbf{B}$$

The last formula is derived as follows. We have

$$\operatorname{rot}(\mathbf{A} \times \mathbf{B}) = \nabla \times (\mathbf{A} \times \mathbf{B}) = \nabla \times (\mathbf{A}_c \times \mathbf{B}) + \nabla \times (\mathbf{A} \times \mathbf{B}_c)$$

Now we must apply the formula for triple vector product (e. g. see [107], Sec. VII.16) and arrange the factors so that ∇ is applied to the factor which is considered variable; thus we obtain

$$\operatorname{rot}(\mathbf{A} \times \mathbf{B}) = \mathbf{A}(\nabla \cdot \mathbf{B}) - (\mathbf{A}\cdot\nabla)\mathbf{B} + (\mathbf{B}\cdot\nabla)\mathbf{A} - \mathbf{B}(\nabla \cdot \mathbf{A}) =$$
$$= \mathbf{A} \operatorname{div}\mathbf{B} - \mathbf{B} \operatorname{div}\mathbf{A} + (\mathbf{B}\cdot\nabla)\mathbf{A} - (\mathbf{A}\cdot\nabla)\mathbf{B}$$

We can also write

$$\operatorname{div}(\mathbf{A} \times \mathbf{B}) = \nabla \cdot (\mathbf{A} \times \mathbf{B}) = \nabla \cdot (\mathbf{A}_c \times \mathbf{B}) + \nabla \cdot (\mathbf{A} \times \mathbf{B}_c)$$

Using the formula for triple scalar product of vectors and arranging the factors appropriately we finally receive

$$\operatorname{div}(\mathbf{A} \times \mathbf{B}) = -\mathbf{A} \cdot (\nabla \times \mathbf{B}) + \mathbf{B} \cdot (\nabla \times \mathbf{A}) = \mathbf{B} \cdot \operatorname{rot}\mathbf{A} - \mathbf{A} \cdot \operatorname{rot}\mathbf{B}$$

3. Integral Formulas. From the Ostrogradsky-Gauss formula (4) we can derive two useful formulas for transforming surface integrals into volume integrals. Let us first put $\mathbf{A}=u\mathbf{i}$ in (4). This results in

$$\oint_{(\sigma)} u\cos(\widehat{\mathbf{n},\mathbf{i}})\,d\sigma=\int_{(\Omega)}\frac{\partial u}{\partial x}d\Omega$$

Multiplying both sides by $\mathbf{i}$, performing analogous operations on the fields $\mathbf{A}=u\mathbf{j}$ and $\mathbf{A}=u\mathbf{k}$ and adding together the results, we obtain (check it up!)

$$\oint_{(\sigma)} u\mathbf{d\sigma}=\oint_{(\Omega)}\operatorname{grad}u\,d\Omega \tag{11}$$

This formula has a simple physical meaning. Let a volume (Ω) be filled with a fluid in the state of rest. Suppose that the action of a field of external forces distributed with volume density $\mathbf{f}$ generates a scalar field of pressure u in the region (Ω). Then we have the equality $\operatorname{grad}u=\mathbf{f}$ (prove it by considering the equilibrium of forces acting upon an infinitesimal cube with sides parallel to the coordinate axes), and therefore the right-hand side of (11) is equal to the resultant of all volume forces applied to (Ω), while the left-hand side of (11) is equal to the sum of forces of pressure acting on the surface (σ) taken with the opposite sign. Thus, in this case equality (11) expresses the equilibrium condition: the resultant of all volume forces is opposite to that of all surface forces, that is the total sum of all forces is equal to zero.

To obtain another formula let us substitute the expression $\mathbf{A}\times\mathbf{i}=A_z\mathbf{j}-A_y\mathbf{k}$ for $\mathbf{A}$ into formula (4). Since we have

$$\operatorname{div}(A_z\mathbf{j}-A_y\mathbf{k})=\frac{\partial A_z}{\partial y}-\frac{\partial A_y}{\partial z}=\operatorname{rot}_x\mathbf{A}\quad\text{and}\quad(\mathbf{A}\times\mathbf{i})\cdot\mathbf{d\sigma}=(\mathbf{d\sigma}\times\mathbf{A})\cdot\mathbf{i}=(\mathbf{d\sigma}\times\mathbf{A})_x$$

formula (4) yields

$$\left(\int_{(\sigma)}\mathbf{d\sigma}\times\mathbf{A}\right)_x=\left(\int_{(\Omega)}\operatorname{rot}\mathbf{A}\,d\Omega\right)_x$$

Now, multiplying both members by $\mathbf{i}$ and adding to the result two similar expressions obtained in the same way for the other two coordinates y and z we finally deduce the formula

$$\int_{(\sigma)}\mathbf{d\sigma}\times\mathbf{A}=\int_{(\Omega)}\operatorname{rot}\mathbf{A}\,d\Omega \tag{12}$$

Dividing both sides of each formula (4), (11) and (12) by Ω (i. e. by the volume of (Ω)) and passing to the limit as the domain (Ω) is contracted to a point we obtain, respectively, invariant defini-

tions (independent of the particular choice of the coordinate system) of the quantities div **A**, grad u and rot **A**. Comparing the resulting formulas with formulas (6), (5) and (7) we can write the following symbolic representation for the Hamiltonian operator:

$$\nabla \ldots = \lim_{\Omega \to 0} \frac{1}{\Omega} \oint_{(\sigma)} \ldots d\boldsymbol{\sigma} = \frac{1}{d\Omega} \oint_{(d\sigma)} \ldots d\boldsymbol{\sigma}$$

The dots designate here the quantity (i.e. a scalar or vector field) the operator ∇ is applied to.

4. Repeated Operations with ∇. After a differential operation has been performed on a field, a new field is obtained on which an operation of that kind can be performed repeatedly. Thus we can consider the *repeated operations* with the symbol ∇. But of course not all possible combinations are meaningful. For instance, the combination div div **A** does not make sense since div **A** is a scalar field whose divergence does not exist because the operation of taking divergence is only applicable to a vector field. Let the reader verify that combining the operations grad, div and rot we obtain exactly five meaningful repeated operations:

$$\text{div grad } u, \quad \text{rot grad } u, \quad \text{grad div } \mathbf{A}, \quad \text{div rot } \mathbf{A}, \quad \text{rot rot } \mathbf{A} \tag{13}$$

Let us first consider the combination rot grad u. It can be written in the form $\nabla \times (\nabla u)$. But, for an "ordinary" vector **a** and an "ordinary" scalar u we always have

$$\mathbf{a} \times (\mathbf{a}u) = \mathbf{0} \tag{14}$$

(why?). Therefore, if we resolve the vector **a** into components along the coordinate axes, substitute this resolution into the left-hand side of (14) and perform the calculations according to the rules of vector algebra, the result will be equal to zero. Now note that the expression $\nabla \times (\nabla u)$ is computed according to the same formal rules as (14), the only difference being that instead of a_x, a_y and a_z we take $\frac{\partial}{\partial x}$, $\frac{\partial}{\partial y}$ and $\frac{\partial}{\partial z}$. Thus, here we must also obtain zero, that is

$$\text{rot grad } u = \mathbf{0} \tag{15}$$

We can similarly prove (check it up!) that for any vector field **A** we always have

$$\text{div rot } \mathbf{A} = 0 \tag{16}$$

The last property has an important consequence. Namely, for an arbitrary vector field **A** we can consider, besides its own vector lines, the vector lines of the field of its rotation rot **A**. As is known (e.g. see [107], Sec. XVI.23), the divergence of a vector field is equal to the density of sources of vector lines of the field.

Hence, formula (16) means that the vector lines of the field rot **A** have neither sources nor sinks, i.e. have neither initial nor terminal points.

The first combination (13) can be written in the form $\nabla \cdot \nabla u = (\nabla \cdot \nabla) u = \nabla^2 u$ where ∇^2 is a scalar differential operator of the second order which is determined by the formula

$$\nabla^2 = \nabla \cdot \nabla = \left(\mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}\right) \cdot \left(\mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}\right) = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.$$

It is known as **Laplace's operator** (or, simply, the **Laplacian**) and is sometimes denoted by the symbol Δ. Thus we have

$$\operatorname{div} \operatorname{grad} u = \nabla^2 u = \Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \tag{17}$$

Finally, let us proceed to consider the last combination (13). Applying the formula for triple vector product (e.g. see [107], Sec. VII.16) and arranging the factors in such a way that the symbols denoting the operations involving ∇ are written in front of the corresponding fields we obtain

$$\operatorname{rot} \operatorname{rot} \mathbf{A} = \nabla \times (\nabla \times \mathbf{A}) = \nabla (\nabla \cdot \mathbf{A}) - (\nabla \cdot \nabla) \mathbf{A} = \operatorname{grad} \operatorname{div} \mathbf{A} - \nabla^2 \mathbf{A} \tag{18}$$

5. Discontinuous Fields. If a field has discontinuities, the application of differential operations may result in appearance of delta-functions and other generalized functions (for elementary facts related to the notion of a generalized function we can refer the reader to [107], Secs. XIV.25 and 27 and XVI.19). Here we shall only consider some types of discontinuity which are most often encountered in physical applications. The systematic theory of generalized functions can be found in [49].

We remind the reader that if a function $f(x)$ has a finite jump at a point $x = a$ (which we shall designate by $[f(a)] \equiv f(a+0) - f(a-0)$) then the differentiation of $f(x)$ results in the expression

$$f'(x) = \varphi(x) + [f(a)]\,\delta(x-a) \tag{19}$$

where $\varphi(x) = f'(x)$ for $x \neq a$ and is an ordinary (i.e. not generalized) function.

Now let us consider a scalar field u possessing a finite jump on the plane $x = 0$:

$$u = \begin{cases} 0 & \text{for } x < 0, \\ f(x, y, z) & \text{for } x > 0 \end{cases}$$

Computing the gradient of u according to formula (1) and using formula (19) we receive

$$\operatorname{grad} u = \varphi(x, y, z) + f(+0, y, z)\,\delta(x)\,\mathbf{i}$$

where

$$\varphi = \begin{cases} 0 & \text{for } x < 0, \\ \operatorname{grad} f & \text{for } x > 0 \end{cases}$$

Note that here the term involving the delta-function only appears in $\frac{\partial u}{\partial x}$ but not in the derivatives $\frac{\partial u}{\partial y}$ and $\frac{\partial u}{\partial z}$. This is accounted for by the fact that when moving along the y-axis or z-axis we encounter no jumps.

If a field u has a finite jump $[u]$ when the variable point passes from the inner side of an oriented surface (S) to its outer side, we can use the fact that the gradient is independent of the particular choice of coordinate axes and thus write

$$\operatorname{grad} u = \varphi(x, y, z) + [u]\,\delta(s)\,\mathbf{n}^0 \tag{20}$$

where $\varphi = \operatorname{grad} f\,((x, y, z)\,\overline{\in}\,(S))$, s is the arc length reckoned from (S) in the direction of the outer normal to the surface and $\mathbf{n}^0$ is the unit vector in this direction which may vary from point to point on the surface (S). (We remind the reader that the symbol "$\in$" means "belongs to" and "$\overline{\in}$" means "does not belong to".) If we have a spatial curvilinear orthogonal coordinate system λ, μ, ν, and (S) is a coordinate surface with equation $\lambda = \lambda_0$, then, instead of $\delta(s)\,\mathbf{n}^0$, we must write $\frac{1}{l_\lambda}\delta(\lambda - \lambda_0)\,\mathbf{e}_\lambda$ in formula (20) where l_λ is a Lamé coefficient (e.g. see [107], Sec. XVI.15) and $\mathbf{e}_\lambda$ is unit tangent vector to the coordinate line λ.

For a vector field $\mathbf{A}$ with finite jump $[\mathbf{A}]$ on a surface (S) we similarly obtain the formula

$$\operatorname{div} \mathbf{A} = \psi(x, y, z) + [A_n]\,\delta(s) \tag{21}$$

where $\psi = \operatorname{div} \mathbf{A}\,((x, y, z)\,\overline{\in}\,(S))$ is an ordinary (not generalized) function. The derivation of (21) is left to the reader. An expression of the form $\sigma\delta(s)$ can be interpreted as the volume density of a mass distributed over (S) with surface density σ, and therefore formula (21) implies that there is a source of vector lines distributed over the surface (S) with density of $[A_n]$ lines per unit area. (We suggest that the reader obtain this result directly by considering the flux of the vector field $\mathbf{A}$ across the surface of an appropriately chosen infinitesimal cylinder.)

Applying the same reasoning we also get the formula

$$\operatorname{rot} \mathbf{A} = \chi(x, y, z) + \mathbf{n} \times [\mathbf{A}]\,\delta(s) \qquad (\chi = \operatorname{rot} \mathbf{A}, (x, y, z)\,\overline{\in}\,(S))$$

Finally, let us consider fields with singular points. For definiteness, we shall suppose that there are singularities only at the origin. As is known (e.g. see [107], Sec. XIV.27), a point func-

tion defined in space and having integrable singularities is not a generalized function but the differentiation of such a function may result in a generalized function. (What has been said applies both to functions of one independent variable and to functions of any number of arguments.) If a function $f(x, y, z)$ has a **polar singularity** at the origin, i.e. it is of the order of r^{-p} $(p > 0)$ for $r \to 0$ (where $r = |\mathbf{r}| = \sqrt{x^2+y^2+z^2}$), then it is integrable if $p < 3$ and nonintegrable if $p \geqslant 3$ (e.g. cf. [107], Sec. XVI.17). On the other hand, it can be readily shown by considering simple examples (for instance, $(x^{-5})' = -5x^{-6}$ and the like) that differentiation increases the order of growth (i.e. the exponent p) by unity, and therefore, when applied to a field having an isolated polar singularity of an order lower than the second, the operation of differentiation results in a singularity of an order lower than the third, etc. Thus, in this case the application of differential operations of the first order does not involve generalized functions.

Isolated polar singularities of order two are of particular interest because they play the same role for spatial fields with singular points as finite jumps for functions of one independent variable. If we take, as the surfaces (σ) entering into formulas (4), (11) and (12), a small sphere of radius r with centre at the origin, then $\sigma = 4\pi r^2$, and the left-hand sides of the formulas have finite limits for $r \to 0$. Therefore, the integrands on the right-hand sides must contain terms with delta-functions. For example, if a field $\mathbf{A}$ has an isolated polar singularity of order two at the origin, then $\operatorname{div}\mathbf{A}$ involves the summand $k\delta(\mathbf{r}) = k\delta(x)\,\delta(y)\,\delta(z)$ where $k = \lim_{r \to 0} \frac{1}{4\pi r^2} \oint_{(\sigma)} \mathbf{A} \cdot d\boldsymbol{\sigma}$; in other words, this field has a point source at the origin which generates k vector lines (see Sec. 4).

For spatial fields with singularity along a line and also for plane fields the same role is played by polar singularities of order one.

§ 2. SPECIAL TYPES OF FIELD

1. Potential Fields. A vector field $\mathbf{A}$ is said to be **potential** if it is the field of gradients of a scalar field, i.e.

$$\mathbf{A} = \operatorname{grad} \varphi \tag{1}$$

The function φ is then called the **potential** (or, more precisely, the **scalar potential**) of the field $\mathbf{A}$.

It should be noted that the potential of $\mathbf{A}$ is sometimes introduced by the relation $\mathbf{A} = -\operatorname{grad}\varphi = \operatorname{grad}(-\varphi)$, and then it differs from (1) by the factor (-1). To distinguish between these cases, besides "potential", the term "**potential function**" is sometimes used but this convention is not commonly recognized.

In what follows we shall use both terms as synonyms, in the sense of formula (1), but the reader should bear in mind this distinction when reading other books in which these notions are encountered and follow the convention assumed there.

Since the gradient of a scalar field is identically equal to zero if and only if the field is constant, we conclude that *the potential of any field is determined to within an arbitrary constant summand.* Choosing this constant term in an appropriate manner we can **normalize** the potential (i.e. get rid of the arbitrariness) by making its value at a chosen point be equal to zero. Most often a potential is normalized in such a way as to make it equal to zero at infinity (provided a finite value can be attributed to it at infinity).

By far not every field is potential. Namely, general formula (1.15) immediately implies that if $\mathbf{A}=\operatorname{grad}\varphi$ then

$$\operatorname{rot}\mathbf{A}=\mathbf{0} \tag{2}$$

Thus, *every potential field is* ***irrotational***. On the other hand, as is known (e.g. see [107], Secs. XVI.27 and XIV.24), *if a field* $\mathbf{A}$ *is defined in a simply connected domain* (G), *condition* (2) *is equivalent to the condition that the integral* $\int \mathbf{A}\cdot d\mathbf{r}$ *is independent of the path of integration* or, which is the same, to the condition

$$\oint_{(L)} \mathbf{A}\cdot d\mathbf{r}=0 \tag{3}$$

for any closed contour (L) lying in (G). Thus, *an irrotational field in a simply connected region is* ***circulation-free (noncirculatory).*** It is also known that condition (3) means that the expression $\mathbf{A}\cdot d\mathbf{r}=A_x dx+A_y dy+A_z dz$ is a total differential. But the equality $\mathbf{A}\cdot d\mathbf{r}=d\varphi$ is equivalent to (1) (why?), and therefore it follows that *every circulation-free field is potential. The converse is also true* (the proof is left to the reader).

If $\mathbf{A}$ is a field of force then the integral $\int_{(L)} \mathbf{A}\cdot d\mathbf{r}$ is equal to the work performed by the field when a particle it acts upon traverses the contour (L). For this case condition (3) has an obvious physical meaning: *a field of force is potential if and only if its work along any closed contour is equal to zero.*

Thus, we see that *for a simply connected domain conditions* (1), (2) *and* (3) *imply each other, i.e. the conditions that a field is potential, irrotational and circulation-free are equivalent.*

2. Irrotational Field in Multiply Connected Domain. Let us consider an irrotational field $\mathbf{A}$ in a *multiply connected* domain (G).

In this case the circulation of the vector **A** over a closed contour is not necessarily equal to zero. The proof of formula (3) only implies that the circulation must be equal to zero if the contour of integration can be contracted to a point without leaving the domain (G). For instance, in Fig. 1 we see a domain which is obtained when two infinite cylinders are removed from the space (this domain is *triply connected*). According to the above, the circulation of the vector field **A** over the contour (L) is equal to zero. As for closed contours that cannot be contracted to a point, we can only assert that *if two contours* (L_1) and (L_1') *can be transformed into each other by means of a continuous deformation without falling outside the domain* (G), *then the circulations of the vector field* **A** *over these contours are equal.* Indeed, if we join the contours with line AD (see Fig. 1) and consider the contour $ABCADEFDA$ shown in dotted line, we find that the latter can be contracted to a point (why?), and therefore

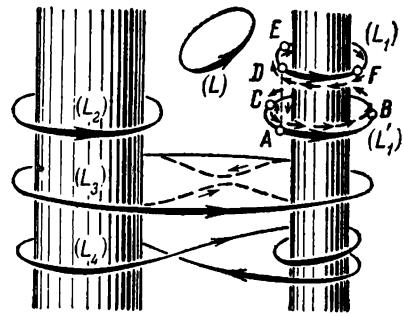


Fig. 1

$$\oint_{ABCADEFDA} \mathbf{A}\cdot d\mathbf{r} = \oint_{ABCA} \mathbf{A}\cdot d\mathbf{r} + \int_{AD} \mathbf{A}\cdot d\mathbf{r} + \int_{DEFD} \mathbf{A}\cdot d\mathbf{r} + \int_{DA} \mathbf{A}\cdot d\mathbf{r} =$$

$$= \oint_{(L_1')} \mathbf{A}\cdot d\mathbf{r} + \int_{AD} \mathbf{A}\cdot d\mathbf{r} - \oint_{(L_1)} \mathbf{A}\cdot d\mathbf{r} + \int_{DA} \mathbf{A}\cdot d\mathbf{r} = 0$$

But the integrals along the joining line mutually cancel out (why?), and consequently we obtain

$$\oint_{(L_1')} \mathbf{A}\cdot d\mathbf{r} - \oint_{(L_1)} \mathbf{A}\cdot d\mathbf{r} = 0, \quad \text{i. e.} \quad \oint_{(L_1')} \mathbf{A}\cdot d\mathbf{r} = \oint_{(L_1)} \mathbf{A}\cdot d\mathbf{r}$$

At the same time, the circulation over the contour (L_2) (Fig. 1) is completely independent, in the general case, of that over the contour (L_1). It is readily seen that in the case of the triply connected domain shown in Fig. 1 the circulations over the contours L_1 and L_2 (we denote them by Γ_1 and Γ_2, respectively) form a complete system of independent circulations in the sense that the circulation over any other closed contour can be expressed in terms of Γ_1 and Γ_2. For instance, the circulation over the contour (L_3) is equal to $\Gamma_1 + \Gamma_2$ (to prove this we should deform (L_3)

as is shown in Fig. 1 by the dotted line), the circulation over (L_4) is equal to $\Gamma_2 - 2\Gamma_1$, etc. For a $(k+1)$-connected domain (also referred to as a domain with **connectivity number** $k+1$) the maximum number of independent circulations is k. In the simplest case of a *doubly connected* domain (for instance, the whole space from which an infinite circular cylinder is cut out, the exterior or the interior of a torus and the like) there is only one independent circulation.

Now suppose that for an irrotational field in a multiply connected domain we construct a potential by using the same formula

$$\varphi(M) = \int\limits_{\smile M_0 M} d\varphi = \int\limits_{\smile M_0 M} \mathbf{A} \cdot d\mathbf{r}$$

(where $\smile M_0 M$ is an arbitrary arc within the domain joining an arbitrary fixed point M_0 to the moving point M) as in the case of a simply connected domain. Then relation (1) obviously holds but $\varphi(M)$ (the potential function) is multiple-valued in the general case. For instance, if we continuously extend the function $\varphi(M)$ along the line (L_1) shown in Fig. 1 and traverse one circuit along that contour, the potential gains the increment $\int\limits_{(L_1)} \mathbf{A} \cdot d\mathbf{r} = \Gamma_1$. The same result is obtained if we describe the contour (L_1'). But if we traverse the contour (L_2) the increment of the potential is equal to Γ_2. For this reason the constants Γ_1 and Γ_2 are sometimes termed the periods of the potential.

Thus, generally speaking, *the potential of an irrotational field in a multiply connected domain is a multiple-valued function*. The field specified by a potential of this type is not considered a potential field. Only when all the periods of the potential are equal to zero or, in other words, when the field is circulation-free, the potential is one-valued and the corresponding field is potential.

3. Solenoidal Fields. *A vector field* **A** *is called* ***solenoidal*** *if it is the field of rotation of another vector field*, that is if

$$\mathbf{A} = \operatorname{rot} \mathbf{\Phi} \tag{4}$$

The field $\mathbf{\Phi}$ is then referred to as the **vector potential** of the field **A**.

The vector potential of a solenoidal field is determined to within the gradient of an arbitrary scalar field. Indeed, if $\mathbf{\Phi}_1 = \mathbf{\Phi} + \operatorname{grad} \psi$ then, by (1.15), we have

$$\operatorname{rot} \mathbf{\Phi}_1 = \operatorname{rot} \mathbf{\Phi} + \operatorname{rot} \operatorname{grad} \psi = \mathbf{A} + \mathbf{0} = \mathbf{A}$$

and the converse can also be readily proved. This property makes it possible to choose the function ψ in an appropriate manner in order to normalize the vector potential.

A field which is solenoidal in a region has no sources of vector lines in this region. In fact, formula (4) and general formula (1.16) imply

$$\operatorname{div} \mathbf{A} = 0 \tag{5}$$

This accounts for the term "solenoidal field" because any magnetic field is a typical example of a vector field without sources, the magnetic field generated by an electric current flowing through a wire helix (a *solenoid*) being of that type.

Now let us show that *a vector field* $\mathbf{A}$ *which is defined in the whole space and has no sources of vector lines is solenoidal.* For this purpose we shall construct its vector potential in the form $\mathbf{\Phi} = P(x, y, z)\,\mathbf{i} + Q(x, y, z)\,\mathbf{j}$. Substituting this expression into (4) and equating the projections of both sides on the coordinate axes we obtain

$$-\frac{\partial Q}{\partial z} = A_x, \quad \frac{\partial P}{\partial z} = A_y, \quad \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} = A_z \tag{6}$$

The first relation (6) determines Q to within an arbitrary function $f(x, y)$. For example, we can put

$$Q(x, y, z) = Q_0(x, y, z) + f(x, y) \quad \text{where} \quad Q_0(x, y, z) = -\int_0^z A_x(x, y, \zeta)\, d\zeta \tag{7}$$

Similarly, the second relation (6) results in $P(x, y, z) = P_0(x, y, z) + f_1(x, y)$ where f_1 is an arbitrary function. Substituting the expressions of Q and P into the third relation (6) and putting $f_1 \equiv 0$ (we can do it because we are only interested in constructing a concrete vector potential) we obtain the equation

$$\frac{\partial Q_0}{\partial x} + \frac{\partial f}{\partial x} - \frac{\partial P_0}{\partial y} = A_z, \quad \text{i. e.} \quad \frac{\partial f}{\partial x} = A_z - \frac{\partial Q_0}{\partial x} + \frac{\partial P_0}{\partial y} \tag{8}$$

which determines f. But the right-hand side of the latter equality is independent of z since, by condition (5), we have

$$\frac{\partial}{\partial z}\left(A_z - \frac{\partial Q_0}{\partial x} + \frac{\partial P_0}{\partial y}\right) = \frac{\partial A_z}{\partial z} - \frac{\partial}{\partial x}\left(\frac{\partial Q_0}{\partial z}\right) + \frac{\partial}{\partial y}\left(\frac{\partial P_0}{\partial z}\right) = \frac{\partial A_z}{\partial z} + \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} = 0$$

Therefore, denoting the right-hand side of (8) by $F(x, y)$ we find $f(x, y) = \int_0^x F(\xi, y)\, d\xi$. Thus the vector potential $\mathbf{\Phi}$ has been constructed.

If a vector field is considered not in the whole space but only in a domain (G), then there appear some difficulties in the application of formula (7). A more detailed investigation of the problem (which we are not going to present here) shows that *for all the vector fields with zero divergence defined in* (G) *to be solenoidal it is necessary and sufficient that it be possible to contract any closed surface of the type of a sphere lying in* (G) *to a point without leaving the domain* (G). (When speaking of "a surface of the type of a sphere" we mean a surface which can be obtained from a sphere by a continuous deformation; for instance, a torus does not belong to this type.) This property is called **acyclicity** of the domain in dimension 2. The property of simple-connectedness (which can be called acyclicity in dimension 1) differs from the above property. For instance, all the convex domains are both acyclic (in dimension 2) and simply connected, a solid bounded by a torus is acyclic but not simply connected, the whole space with a sphere cut out of it is simply connected but not acyclic and, finally, the space with a torus removed from it is neither simply connected nor acyclic. It can be proved that a domain in the three-dimensional space is acyclic if and only if it is finite and has a connected (i. e. consisting of one component) boundary or is infinite and all the connected components of its boundary recede to infinity.

The flux of a solenoidal field Φ *through any closed surface* (σ) *is equal to zero.* If the field in question is defined everywhere inside (σ), this assertion follows from the Ostrogradsky-Gauss formula (1.4). To prove the assertion for the general case, we break up (σ) into small areas $(d\sigma)$ and choose the orientations of their contours in such a way that they are coherent with the orientation of (σ). Then the sum of the circulations $d\Gamma$ of the vector field Φ over all the contours is equal to zero (why?). But $d\Gamma = \mathrm{rot}_n \Phi \, d\sigma = A_n \, d\sigma$ and hence $\oint\limits_{(\sigma)} A_n \, d\sigma = \oint d\Gamma = 0$ which is what we set out to prove.

Conversely, *if*

$$\oint\limits_{(\sigma)} A_n \, d\sigma = 0 \quad \textit{for any closed surface } (\sigma) \tag{9}$$

in a domain (G), *the field* $\mathbf{A}$ *is solenoidal in this domain.* If the domain (G) is the whole space, then this assertion follows from (5) and the fact proved in the paragraph after formula (5), relation (5) being implied in this case by (9) and the invariant definition of divergence. For an arbitrary domain (G) the proof is more complicated and we shall not present it here.

Now we can explain the significance of the above acyclicity condition. If this condition is fulfilled and, besides, we have

div $\mathbf{A}=0$ everywhere in (G), then for any closed surface (σ) lying in (G) the field $\mathbf{A}$ is defined everywhere inside (σ), and therefore we can apply the Ostrogradsky-Gauss formula to (σ) which results in (9). Now suppose that the domain (G) is not acyclic; for instance, let (G) be obtained from the whole space by removing two spheres (K_1) and (K_2) from it. Consider two surfaces (σ_1) and (σ_2) lying in (G) such that each (σ_i) $(i=1, 2)$ contains the sphere (K_i) in its interior and does not contain the points of the other sphere. Then, generally speaking, condition (5) does not necessarily imply that

$$\int\limits_{(\sigma_1)} A_n\,d\sigma=0 \quad \text{and} \quad \int\limits_{(\sigma_2)} A_n\,d\sigma=0 \tag{10}$$

At the same time, if both conditions (5) and (10) hold they imply (9), that is in this case the field $\mathbf{A}$ is solenoidal.

In conclusion we note that a field $\mathbf{A}$ may be *simultaneously potential and solenoidal*; then it has neither vortices nor sources of vector lines in the given region. Therefore we have $\mathbf{A}=\operatorname{grad}\varphi$, and, since div $\mathbf{A}=0$, this means that

$$\operatorname{div}\operatorname{grad}\varphi=0, \text{ i. e. (see Sec. 1.4) } \nabla^2\varphi=0 \tag{11}$$

Equation (11) is known as the **Laplace equation** and its solutions are called **harmonic functions** (this term must not be confused with the notion of a harmonic, i. e. sinusoidal, functional relation). Thus, *for a potential field to be solenoidal it is necessary that the potential of the field be a harmonic function*. For a field considered in the whole space and for a more general case when a field is defined in an acyclic domain this condition is both necessary and sufficient.

The equation $\nabla^2\varphi=0$ had been known to L. Euler but was named after P. S. Laplace who investigated this equation in 1782 and applied it to the theory of gravitation.

4. Examples. Let us consider a centrally symmetric space vector field determined by the formula

$$\mathbf{A}=f(r)\,\mathbf{r}^0=\frac{f(r)}{r}\,\mathbf{r} \quad \left(\mathbf{r}=x\mathbf{i}+y\mathbf{j}+z\mathbf{k},\ \ \mathbf{r}^0=\frac{\mathbf{r}}{r}\right) \tag{12}$$

From the definition of divergence we can easily deduce (e. g. see [107], Sec. XVI.23) the formula $\operatorname{div}\mathbf{A}=\frac{1}{r^2}\frac{d}{dr}(r^2f(r))$. In particular, for a centrally symmetric field having no sources other than at the origin we can write $r^2f(r)=\text{const}$, whence

$$f(r)=\frac{C}{r^2}, \quad \mathbf{A}=\frac{C}{r^3}\,\mathbf{r} \tag{13}$$

where C is a constant. This is known as **Coulomb's law** named after C. A. Coulomb (1736-1806), the celebrated French physicist, who

was the first to discover the application of this law to the theory of electric and magnetic interactions. The flux of this field through the surface of any sphere of radius r with centre at the origin is equal to $\frac{C}{r^2} \cdot 4\pi r^2 = 4\pi C$, i. e. does not depend on r. In other words, in this case there are $4\pi C$ vector lines starting from the origin and extending to infinity. Let the reader find out where are the initial and terminal points (they can also lie at infinity) of the vector lines in the case when the function $f(r)$ increases slower or faster than r^{-2} for $r \to 0$.

A point source of intensity $Q = 4\pi C$ is characterized by the density $Q\delta(\mathbf{r}) = Q\delta(x)\,\delta(y)\,\delta(z)$, and thus we arrive at the important formula

$$\operatorname{div}\left(\frac{Q}{4\pi r^2}\,\mathbf{r}^0\right) = Q\delta(\mathbf{r})$$

In the case of a plane centrally symmetric field of type (12) for which $\mathbf{r} = x\mathbf{i} + y\mathbf{j}$ we similarly conclude that if there are no sources other than at the origin then $f(r) = \frac{C}{r}$, $Q = 2\pi C$ and

$$\operatorname{div}\left(\frac{Q}{2\pi(x^2+y^2)}(x\mathbf{i}+y\mathbf{j})\right) = Q\delta(x\mathbf{i}+y\mathbf{j}) = Q\delta(x)\,\delta(y)$$

This field can be interpreted as being generated by a point source of vector lines with strength Q located in the plane or by a source distributed with density of Q vector lines per unit length along the z-axis in space.

Now consider a field generated by a superposition of a source and a sink of the same strength (intensity) placed infinitely close to each other. It should be noted that if their intensities are finite, the corresponding fields mutually cancel out. Therefore, for the resultant field to be nonzero these intensities must be infinitely large and such that the product of the intensity by the distance between the source and sink is finite (we denote this quantity by m). The combination of this type is referred to as a **dipole.** A dipole is completely characterized by the quantity m and the axis (the *axis of the dipole* which we denote by l) passing through the source and sink in the direction from the latter to the former. If $\mathbf{l}^0$ is unit vector of l, the vector $\mathbf{m} = m\mathbf{l}^0$ is called the **dipole moment** *) **(or polarization).**

To derive the expression for the vector field of a dipole, consider Fig. 2. At every point M, for a sufficiently small h, we have

$$\mathbf{A} = \frac{Q}{4\pi r_1^3}\mathbf{r}_1 - \frac{Q}{4\pi r^3}\mathbf{r} = Qh\,\frac{\dfrac{\mathbf{r}_1}{4\pi r_1^3} - \dfrac{\mathbf{r}}{4\pi r^3}}{h} = m\frac{d}{dl}\left(\frac{\mathbf{r}}{4\pi r^3}\right) \qquad (14)$$

*) The scalar quantity m is sometimes also referred to as *dipole moment*.— *Tr.*

Simplifying the right-hand side of (14) we obtain (check it up!) the formula

$$\mathbf{A}=\frac{m}{4\pi}\left(\frac{1}{r^3}\frac{d\mathbf{r}}{dl}-\frac{3\mathbf{r}}{r^4}\frac{dr}{dl}\right)=\frac{m}{4\pi r^3}\left(3\frac{\mathbf{r}}{r}\cos\alpha-\mathbf{l}^0\right)$$

The technique used here for determining field (14) is not only applicable to fields of type (13) but also to any arbitrary fields.

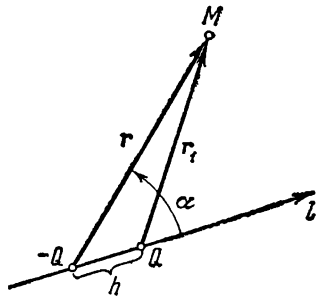


Fig. 2

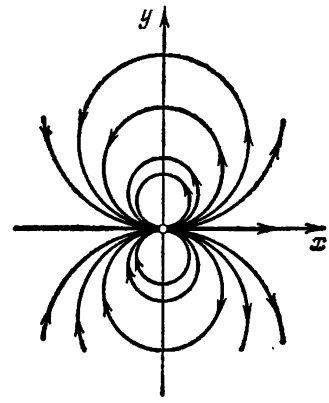


Fig. 3

In the case of plane fields we similarly obtain the following expression for the field of a dipole:

$$\mathbf{A}=\frac{m}{2\pi r^2}\left(2\frac{\mathbf{r}}{r}\cos\alpha-\mathbf{l}^0\right)$$

For simplicity, let the axis l coincide with the x-axis. Then in the Cartesian coordinates x, y this field is written in the form

$$\mathbf{A}=\frac{m}{2\pi(x^2+y^2)}\left(2\frac{x\mathbf{i}+y\mathbf{j}}{x^2+y^2}x-\mathbf{i}\right)=\frac{m[(x^2-y^2)\mathbf{i}+2xy\mathbf{j}]}{2\pi(x^2+y^2)^2}$$

The vector lines of the field are specified by the differential equation $\frac{dx}{A_x}=\frac{dy}{A_y}$ whose integration yields (check it up!) $x^2+y^2=Cy$ where C is an arbitrary constant. These vector lines are shown in Fig. 3.

5. Newtonian Potential. It is readily seen that the potential of field (12) is of the form $\varphi(r)=\int f(r)\,dr$. In particular, the Coulomb field (13) has the potential $-\frac{C}{r}=-\frac{Q}{4\pi r}$ where Q is the intensity of the source of vector lines located at the point $r=0$. Now suppose that a source of vector lines of an irrotational field $\mathbf{A}$ is

distributed in space with volume density $f(\mathbf{r}) = f(x, y, z)$. This source can be regarded as a superposition of infinitesimal sources each of which lies at a certain point (x_0, y_0, z_0) and is characterized by its intensity $f(\mathbf{r}_0)\,d\Omega_0$ $(d\Omega_0 = dx_0\,dy_0\,dz_0)$. The value of the potential of such an infinitesimal source at the point with radius vector $\mathbf{r}$ is equal to

$$d\varphi(\mathbf{r}) = -\frac{f(\mathbf{r}_0)\,d\Omega_0}{4\pi|\mathbf{r}-\mathbf{r}_0|}$$

Summing up the elementary potentials we obtain the potential of the vector field $\mathbf{A}$:

$$\varphi(\mathbf{r}) = -\frac{1}{4\pi}\int\frac{f(\mathbf{r}_0)}{|\mathbf{r}-\mathbf{r}_0|}\,d\Omega_0 = \\ = -\frac{1}{4\pi}\iiint\frac{f(x_0, y_0, z_0)}{\sqrt{(x-x_0)^2+(y-y_0)^2+(z-z_0)^2}}\,dx_0\,dy_0\,dz_0 \qquad (15)$$

The integral is taken here over the whole space or, which is the same, over the region occupied by the source, i. e. over the domain specified by the relation $f \neq 0$. Expression (15) is called the **Newtonian potential** (corresponding to the source density f).

The above deduction of formula (15) is an example of a standard technique based on the notion of an *influence function* (e. g. see [107], Secs. XIV.26 and XVI.19): the transition from a given density of sources of an irrotational vector field to its potential is determined by the linear operator corresponding to the influence function $-\frac{1}{4\pi|\mathbf{r}-\mathbf{r}_0|}$ where $\mathbf{r}_0$ is the radius vector of the point of application of action and $\mathbf{r}$ is that of the point of observation.

The result thus obtained has a clear physical meaning. Namely, a point charge q placed at the origin produces in vacuum the electric field $\mathbf{E} = k\frac{q}{r^2}\mathbf{r}^0$ where the coefficient k depends on the choice of the system of units. This field has at the origin a source of vector lines of intensity $Q = 4\pi kq$ whose potential is $-k\frac{q}{r}$. Therefore, expression (15) can be regarded as the potential of the electric field generated by electric charges distributed in space with density $\frac{1}{4\pi k}f(\mathbf{r})$. For a gravitational field generated by masses distributed in space we obtain the same result but with the opposite sign.

For simplicity, let f be a **finite** *) function (i. e. identically equal

*) The notion of a *finite* function, which vanishes outside some finite, i.e. bounded, region, must not be confused with the notion of a *bounded* function whose range is contained in some finite interval (although we sometimes say that a function is finite when it does not take on infinite values). Finite functions are also called "**functions of finite** (or **compact**) **support**".— *Tr.*

to zero outside a bounded region) which assumes finite values. Then the integrand in integral (15) has only one singularity at the point $\mathbf{r}_0 = \mathbf{r}$. But if we pass to spherical coordinates with centre at the point (x, y, z), the singularity disappears. If the function f has some other singularities or the domain where $f \neq 0$ extends to infinity, we must additionally impose some conditions for convergence of the integral; for instance, in the latter case for the integral to be convergent it is sufficient that $|f(x, y, z)|$ decrease for $r \to \infty$ faster than r^{-2} (why?).

The above result admits of another interpretation. We have $f(\mathbf{r}) = \operatorname{div} \mathbf{A}(\mathbf{r}) = \operatorname{div} \operatorname{grad} \varphi(\mathbf{r}) = \nabla^2 \varphi$ (see Sec. 1.4), and therefore the Newtonian potential (15) is a solution of the equation

$$\nabla^2 \varphi = f(x, y, z) \tag{16}$$

This is a (three-dimensional) **Poisson equation** (derived by S. D. Poisson in 1813) which can be regarded as a nonhomogeneous equation corresponding to the Laplace (homogeneous) equation (11). It can easily be shown that $\varphi(\mathbf{r}) \to 0$ for $r \to \infty$ if the assumptions of the foregoing paragraph hold. It can also be proved (but we do not give the proof here) that equation (16) has a unique solution satisfying this condition at infinity, namely the one determined by integral (15).

The simple case when $f \equiv 0$ is an example showing that the condition $\lim\limits_{r \to \infty} \varphi(\mathbf{r}) = 0$ is essential for equation (16) to have a unique solution. In fact, if we do not impose this condition, then, besides the "basic" solution $\varphi \equiv 0$, there appear not only solutions of the form $\varphi = \text{const}$ which are of no importance for the potential theory but also "nontrivial" solutions of the form $\varphi = ax + by + cz$ (which are the potentials of dipoles located at infinity) and other solutions whose rate of growth is still greater.

For plane fields, instead of (15), we similarly arrive at the **logarithmic potential**

$$\varphi(\mathbf{r}) = \frac{1}{2\pi} \int f(r_0) \ln |\mathbf{r} - \mathbf{r}_0| \, d\sigma_0 =$$

$$= \frac{1}{2\pi} \int f(x_0, y_0) \ln \sqrt{(x - x_0)^2 + (y - y_0)^2} \, dx_0 \, dy_0$$

which satisfies the *two-dimensional Poisson equation* $\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} = f$.

If the function f is finite and assumes finite values, then, in the general case, this potential approaches infinity as $r \to \infty$ but its derivatives tend to zero. This condition determines the solution of the two-dimensional Poisson equation to within an arbitrary constant summand.

6. Constructing Vector Field with Given Rotation and Divergence. In the theory of general vector fields we encounter the

problem of determining a vector field Φ when the fields

$$\operatorname{rot}\Phi = \mathbf{A}(\mathbf{r}) \quad \text{and} \quad \operatorname{div}\Phi = \alpha(\mathbf{r}) \tag{17}$$

are given. We shall consider the fields Φ, $\mathbf{A}$ and α in the whole space and suppose that $\mathbf{A}$ and α are finite and tend to zero sufficiently fast for $r \to \infty$. Besides, we shall impose the condition $\operatorname{div}\mathbf{A} = 0$ which is necessary in this case (see Sec. 3). It turns out that *if we additionally introduce the requirement*

$$\lim_{r \to 0} \Phi(\mathbf{r}) = 0 \tag{18}$$

this problem is solvable and the solution is unique.

To construct the solution, let us take the rotations of both sides of the first relation (17) and apply formula (1.18). This results in

$$\nabla^2\Phi = \operatorname{grad}\operatorname{div}\Phi - \operatorname{rot}\operatorname{rot}\Phi = \operatorname{grad}\alpha - \operatorname{rot}\mathbf{A} \tag{19}$$

The right-hand side being given, we thus obtain a Poisson equation of form (16) for the vector field Φ. But we can project both sides of this equation on the coordinate axes, and therefore the solution $\Phi(\mathbf{r})$ of equation (19) is expressible in the form of a Newtonian potential (think about it!) and satisfies condition (18).

To show that the field Φ thus found from equation (19) satisfies equations (17) as well, we rewrite (19) in the form

$$\operatorname{grad}(\operatorname{div}\Phi - \alpha) = \operatorname{rot}(\operatorname{rot}\Phi - \mathbf{A}) \tag{20}$$

Taking the divergence of both sides of this equation we see that $\nabla^2(\operatorname{div}\Phi - \alpha) = 0$ and therefore, by condition (18), we conclude that $\operatorname{div}\Phi - \alpha \equiv 0$, that is the second relation (17) is fulfilled. If we now take the rotations of both sides of (20) and use the relation $\operatorname{div}(\operatorname{rot}\Phi - \mathbf{A}) = 0 - \operatorname{div}\mathbf{A} = 0$, we similarly prove that the first equation (17) is also fulfilled.

The solution we have constructed is unique. Indeed, the difference of two fields satisfying equations (20) is a solution of the corresponding homogeneous equations (in which $\mathbf{A} \equiv 0$ and $\alpha \equiv 0$), and therefore the uniqueness is implied by (19) (why?).

In particular, it follows that *every finite vector field* $\mathbf{A}$ *(or a field* $\mathbf{A}$ *that tends to zero sufficiently fast when the variable point travels to infinity) is representable in the form of a sum of a potential field and a solenoidal field*:

$$\mathbf{A} = \operatorname{grad}\varphi + \operatorname{rot}\Phi \tag{21}$$

If the additional conditions $\operatorname{div}\Phi = 0$, $\lim_{r \to \infty} \varphi = 0$ *and* $\lim_{r \to \infty} \Phi = 0$ *hold the representation is unique.*

In fact, taking the divergence of both sides of (21) we receive the equation $\nabla^2\varphi = \operatorname{div}\mathbf{A}$ from which, under the condition $\lim_{r \to \infty} \varphi = 0$,

the function φ is determined uniquely. Then we rewrite the conditions for $\mathbf{\Phi}$ in the form

$$\operatorname{rot}\mathbf{\Phi} = \mathbf{A} - \operatorname{grad}\varphi, \quad \operatorname{div}\mathbf{\Phi} = 0 \quad \text{and} \quad \lim_{r \to \infty} \mathbf{\Phi} = \mathbf{0}$$

which enables us to apply the result established above. Thus the field $\mathbf{\Phi}$ is also determined uniquely. It should be noted that, by (19), the equation for $\mathbf{\Phi}$ is of the form

$$\nabla^2\mathbf{\Phi} = \operatorname{grad} 0 - \operatorname{rot}(\mathbf{A} - \operatorname{grad}\varphi) = -\operatorname{rot}\mathbf{A}$$

and hence the field $\mathbf{\Phi}$, like φ, is expressible in the form of a Newtonian potential.

CHAPTER II

The Theory of Analytic Functions

In this chapter we shall study complex functions of a complex variable. The reader is supposed to be familiar with basic properties of complex numbers including the simplest algebraic and transcendental operations on them and also with the notion of a power series with complex terms (e. g. see [107], § VIII.1, Secs. VIII.8 and 11, and Sec. XVII.14). The theory of analytic functions is an extensive branch of mathematics having many applications. Some of its divisions, for instance, the theory of conformal mappings, are directly related to the theory of plane fields. It is also applicable to computing integrals, investigating series, solving differential equations, etc. in various problems connected with analytic methods of mathematics.

Here we shall only present some basic facts of the theory and special problems which are important for applications. For more detail we refer the reader to [41], [42], [81], [88], [93], [105], [120], [123] and [126].

§ 1. DIFFERENTIATION AND MAPPINGS

1. Derivative. Consider a single-valued function $w=f(z)$ of a complex variable z which is defined in the complex plane or in its part and assumes complex values. As is known (e. g. see [107], Sec. VIII.11), the derivative of $f(z)$ is defined by the formula

$$\frac{dw}{dz}=f'(z)=\lim_{\Delta z\to 0}\frac{\Delta w}{\Delta z}=\lim_{\Delta z\to 0}\frac{f(z+\Delta z)-f(z)}{\Delta z} \tag{1}$$

where the limit must be the same for all the possible ways in which Δz approaches zero. Formula (1) is known from the theory of real functions, and all the basic properties of differentiation remain true for the case of a complex function.

A function which is continuous in a domain of the complex plane and possesses a finite derivative at each point of the domain is said to be **analytic** *in this domain.* Roughly speaking, the condition of

analyticity of a function $f(z)$ not only means that the function must not have points and lines of discontinuity but also that the values of $f(z)$ are obtained from z by operations (algebraic operations, operation of forming sums of infinite series and other operations involving passages to limits) in which z enters as a complex aggregate without separation of its real and imaginary parts. For instance, such is the function $w=z^2$ although it can be written in the form $w=x^2+i2xy-y^2$ if we put $z=x+iy$. On the other hand, the functions $s=x^2-iy^2$ and $t=z^*=x-iy$ dependent on z are not analytic in z. Indeed, for every given value of z we can find the corresponding values of $x=\operatorname{Re} z$ and $y=\operatorname{Im} z$ and compute $s=s(x, y)$ and $t=t(x, y)$ which thus are functions of z, but $s(x, y)$ and $t(x, y)$ cannot be expressed in the form of algebraic or analytic expressions of the types $s(z)$ and $t(z)$.

It should be noted that from the point of view of the above definition some functions $\left(\text{for instance, } \frac{1}{z}\right)$ cannot be regarded as being analytic throughout the complex plane z. The latter function is analytic in the domain obtained by deleting the point $z=0$ at which it is discontinuous and thus not analytic. A point of this kind is referred to as a **singular point** of the analytic function. In such cases we sometimes use another terminology: the function $\frac{1}{z}$ is said to be analytic in the whole z-plane, and regular everywhere except the point $z=0$ which is referred to as a **singularity** (singular point) of the function.

2. Cauchy-Riemann Equations. Suppose we are given a function

$$w=f(z) \tag{2}$$

Let us introduce the notation

$x=\operatorname{Re} z$, $y=\operatorname{Im} z$, $u=\operatorname{Re} w$, $v=\operatorname{Im} w$, i.e. $z=x+iy$, $w=u+iv$

Then, for given x and y, we can determine z and find, by means of formula (2), the value of w which specifies u and v. Hence, we can write

$$u=u(x, y), \quad v=v(x, y) \tag{3}$$

Conversely, if functions (3) are given, they uniquely determine function (2). Thus, the specification of a complex function (2) of a complex argument is equivalent to the specification of two real functions (3) of two real variables. As an exercise, show that the relation $w=e^z$ is equivalent to the system of two equalities

$$u=e^x\cos y, \quad v=e^x\sin y \tag{4}$$

If it is required that function (2) be analytic, then two functions (3) must satisfy certain relations which we are now going to

derive. Note that according to (3) we have

$$\frac{dw}{dz}=\frac{d(u+iv)}{d(x+iy)}=\frac{(u+iv)'_x\,dx+(u+iv)'_y\,dy}{dx+i\,dy} \tag{5}$$

The quantity $dz=dx+i\,dy$ being arbitrary, the differentials dx and dy can be in an arbitrary ratio t. Dividing the numerator and denominator of the right member of (5) by dy and then performing the division of the numerator by the denominator we receive

$$\frac{dw}{dz}=\frac{(u+iv)'_x\,t+(u+iv)'_y}{t+i}=(u+iv)'_x+\frac{(u+iv)'_y-i\,(u+iv)'_x}{t+i}$$

For the derivative (understood in the sense of definition (1)) to exist it is necessary and sufficient that the expression on the right-hand side be independent of t, that is $(u+iv)'_y-i\,(u+iv)'_x=0$ which is equivalent to $u'_y+iv'_y=iu'_x-v'_x$. Now, equating the real and imaginary parts, we arrive at the **Cauchy-Riemann equations**

$$\frac{\partial u}{\partial x}=\frac{\partial v}{\partial y},\quad \frac{\partial v}{\partial x}=-\frac{\partial u}{\partial y} \tag{6}$$

which are necessary and sufficient for function (2) to be analytic. (Verify that these equations hold for function (4).)

Relations (6) can be rewritten in another form. Consider two arbitrary axes n and m starting from a point (x, y) in the plane and directed so that n is obtained from m by counterclockwise rotation through 90°. Then, using the formula for the directional derivative (e.g. see [107], Sec. XII.1) and equations (6) we obtain (check it up!)

$$\begin{aligned}\frac{\partial v}{\partial n}&=\frac{\partial v}{\partial x}\cos\widehat{(n,x)}+\frac{\partial v}{\partial y}\cos\widehat{(n,y)}=-\frac{\partial u}{\partial y}\cos\widehat{[(m,x)+90^\circ]}+\\&+\frac{\partial u}{\partial x}\cos\widehat{(m,x)}=\frac{\partial u}{\partial x}\cos\widehat{(m,x)}+\frac{\partial u}{\partial y}\sin\widehat{(m,x)}=\\&=\frac{\partial u}{\partial x}\cos\widehat{(m,x)}+\frac{\partial u}{\partial y}\cos\widehat{(m,y)}=\frac{\partial u}{\partial m}\end{aligned} \tag{7}$$

Conversely, equations (6) are readily derived from (7) (how?).

3. **Conjugate Harmonic Functions.** The function v can be eliminated from equations (6). To this end, let us differentiate the first equation with respect to x and the second with respect to y; then the right member of the former becomes equal to the left member of the latter (why?). Equating the other members we deduce

$$\frac{\partial^2 u}{\partial x^2}=-\frac{\partial^2 u}{\partial y^2},\quad \text{i.e.}\quad \frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0$$

Consequently (cf. Sec. I.2.3), $u(x, y)$ is a *harmonic function*. Similarly (let the reader verify it), eliminating the function u from equations (6) we conclude that the function $v(x, y)$ is also harmonic. Two functions (u and v) satisfying the Cauchy-Riemann equations are called **conjugate harmonic functions**. Thus, *the real and imaginary parts of an analytic function are conjugate harmonic functions.*

When considering two conjugate harmonic functions we should distinguish between the first (u) and the second (v) functions. Let the reader show that if the order of the functions is reversed then for the functions to be conjugate one of them must be multiplied by -1.

For any given harmonic function it is possible to construct another harmonic function conjugate to it. For example, let a harmonic function $v(x, y)$ be given. Then equations (6) from which the function $u(x, y)$ must be found can be rewritten in the form of one relation

$$du \equiv \frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial y}dy = \frac{\partial v}{\partial y}dx - \frac{\partial v}{\partial x}dy \tag{8}$$

We have thus arrived at the problem of reconstructing a function from its total differential (e.g. see [107], Sec. XIV.24). The conditions for solvability of the problem for the given right-hand member of (8) are equivalent to the requirement that v be a harmonic function (check it up!). Hence, the problem is in fact solvable and its solution is determined, to within an arbitrary constant summand, by the formula

$$u(x, y) = \int_{(x_0, y_0)}^{(x, y)} du + \text{const} = \int_{(x_0, y_0)}^{(x, y)} \left(\frac{\partial v}{\partial y}dx - \frac{\partial v}{\partial x}dy\right) + \text{const}$$

If the domain in which this construction is made is multiply connected the function u specified by the last formula may be multiple-valued. Then it gains some constant increments when the variable point traverses contours enveloping the "lacunas" (cf. Sec. I.2.2).

4. Geometric Interpretation of Derivative. As is known (e.g. see [107], Sec. VIII.11), a function of type (2) determines a **mapping** *) of the complex z-plane (or of a domain of the z-plane if the values of z are only taken in it) into the w-plane. This mapping can be expressed by formulas (3) involving real quantities. If

*) Generally, in modern mathematics the terms "*function*" and "*mapping*" (and also "*transformation*" and "*operator*") are often used as synonyms. Accordingly, when one object is obtained from another by a mapping (transformation, etc.), the latter is referred to as the *image* or *transform* and the former is termed the *preimage, inverse image, inverse transform* or *original.—Tr.*

function (2) is one-valued (we shall see later on that the theory of analytic functions also deals with multiple-valued functions; in what follows these cases will be carefully stipulated) the mapping is also *one-valued* (*single-valued*), i.e. with each point z only a single point w is associated. If to different values of z there correspond different values of w, or, in other words, if the inverse of (2) is also a one-valued function, the mapping determined by formula (2) is said to be **univalent** or **one-sheeted**. This term means that under the mapping (2) the points w cover the w-plane (or a region of the w-plane) only once when the point z runs over the z-plane (or over the corresponding domain of the z-plane). (If otherwise, the mapping is called **multivalent** or **many-sheeted** or is said to be **many-to-one**; it can be 2-valent, 3-valent, and, generally, n-valent or even infinitely-valent.) Univalent mappings are also said to be **one-to-one** (e.g. see [107], Sec. XI.13). These mappings are the simplest because, under a mapping of that type, to each z there corresponds exactly one w and, conversely, to each w there corresponds exactly one z. (Here we always mean the whole z-plane and w-plane or the corresponding domains lying in them.)

If function (2) is analytic we say that the corresponding mapping is also **analytic**. Let us consider a mapping of this kind in a small neighbourhood of a fixed point z_0 under the additional assumption

$$f'(z_0) \neq 0 \tag{9}$$

Introducing the notation

$$z - z_0 = \Delta z, \quad w_0 = f(z_0), \quad w - w_0 = \Delta w, \quad |f'(z_0)| = k_0, \quad \arg f'(z_0) = \alpha_0$$

we can write, to within infinitesimals of higher order, the relation

$$\Delta w \approx dw = f'(z_0)\,\Delta z, \quad \text{i.e.} \quad |\Delta w| = k_0 |\Delta z| \quad \text{and} \quad \operatorname{Arg} \Delta w = \operatorname{Arg} \Delta z + \alpha_0$$

Imagine now that the planes z and w are made coincident. Then under the mapping in question every small vector Δz with origin at the point z_0 goes into a vector Δw with origin at w_0 and thus Δz undergoes a *translation*, a *k_0-fold magnification* and a *rotation* through an angle α_0. Consequently, to within infinitesimals of higher order, the small neighbourhood of the point z_0 undergoes the three indicated transformations, that is **translation, uniform magnification** *) and **rotation**, *the values of the modulus and argument of the derivative at the point z_0 being, respectively, the* **magnification ratio** *and* **angle of rotation.**

---

*) Here the word *magnification* is understood in a general sense and can correspond to *stretching* if $|f'(z_0)| > 1$ or *shrinking* if $|f'(z_0)| < 1$ (or neither if $|f'(z_0)| = 1$).—*Tr.*

Thus, if in a domain (G) of the z-plane we have $f'(z) \neq 0$ at each point then, to within infinitesimals of higher order, the image of every small figure lying in it is geometrically similar to the preimage. In particular, we see that under an analytic mapping a small circle goes into a circle (but not into an ellipse as it can be in the case of a general mapping (e.g. see [107], Sec. XI.14)), and the angles between intersecting lines are preserved. It also follows that an analytic mapping preserves the orientation of the plane, that is if a small closed contour is described in the z-plane the point w traverses its image in the same direction. Conversely, it can be proved that if a mapping of the plane preserves the similarity of infinitesimal geometric figures (or only the angles between intersecting lines) and the orientation of the plane, the mapping is analytic.

The angle-preserving property implies, in particular, that the lines $u(x, y) = C_1$ and $v(x, y) = C_2$ form two mutually orthogonal families of curves in the xy-plane (why?). Therefore, taking various analytic functions $f(z)$ we can obtain the corresponding curvilinear orthogonal coordinate systems in the plane (see the examples below). The reader is invited to establish the orthogonality on the basis of equality (7).

At the points where condition (9) is violated the structure of mapping (2) differs from that described above; we shall consider this question in Sec. 2.6. But if $f(z) \neq \text{const}$ the presence of a finite number of such points does not influence the validity of our general conclusions concerning the character of an analytic mapping.

5. Conformal Mappings. A one-to-one mapping of a plane region which preserves the similarity of infinitesimal figures and the orientation of the plane is termed **conformal** *) (or, more precisely, is called a **conformal mapping of the first kind**). If the similarity is preserved but the orientation changes to the opposite, the mapping is called a **conformal mapping of the second kind.** It follows from Sec. 4 that for a mapping determined by function (2) to be conformal it is necessary and sufficient that this function and its inverse function be one-valued and analytic in the corresponding domains. We have $\frac{dz}{dw} = \left(\frac{dw}{dz}\right)^{-1}$ and therefore the condition of analyticity of the inverse function is equivalent to the relation $f'(z) \neq 0$.

Generally speaking, only infinitesimal figures go into similar geometric figures under a conformal mapping whereas the form of

*) If a domain (G) of the z-plane is conformally mapped on a domain (H) of the w-plane $(w = f(z))$, the domain (H) is said to be a **conformal image** of (G).—*Tr.*

finite figures may considerably change. For instance, in Fig. 4 we see the image of the square $ABCD$ in the z-plane which is divided into 16 small squares. The image is the curvilinear figure $A'B'C'D'$ in the w-plane, the corresponding curves intersecting at right angles. Although every small part of the z-plane undergoes a uniform magnification and rotation, the magnification ratio and angle of rotation vary from point to point which results in the distortion of the shape of the entire domain.

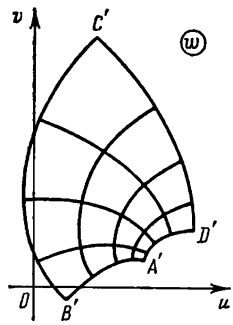


Fig. 4

Now we can discuss the structure of a conformal mapping of the second kind of the z-plane into the w-plane. If we additionally perform the **reflection transformation** $w_1 = w^*$ the resultant transformation is a conformal mapping of the first kind (why?), and thus $w_1 = f(z)$ which implies $w = [f(z)]^*$ where $f(z)$ is an analytic function.

Conformal mappings are widely applied in the theory of plane fields. Examples of conformal mappings will be given below.

6. Linear Mappings. Linear mappings are defined by the formula

$$w = az + b \tag{10}$$

where a and b are complex numbers. They form the simplest class of mappings. We have $\frac{dw}{dz} = a = \text{const}$ and therefore all the small parts of the plane undergo the same $|a|$-fold magnification and rotation through the same angle $\arg a$. This means that the whole plane itself is subjected to the $|a|$-fold magnification and rotation by the angle $\arg a$. If we consider the z-plane and the w-plane to be coincident and put $b = 0$, the point $z = 0$ goes into itself under the mapping. If $b \neq 0$ the **fixed point** of the mapping is determined by the equation $az + b = z$.

Hence, a linear conformal mapping is also linear in the sense of the general theory of mappings (e.g. see [107], Sec. XI.6). It should be noted that a general linear mapping of a plane is not necessarily conformal. For instance, a stretching along an axis or a shift changes angles between intersecting lines and thus is not a conformal mapping.

Consider some special cases. If $a = e^{i\alpha}$ where α is real, the mapping determined by formula (10) is the rotation of the complex plane through the angle α. For example, if $a = i = e^{i\frac{\pi}{2}}$, the plane is turned in the positive direction by 90°. If a is real and $a > 0$, mapping (10) results in an a-fold uniform magnification. If $a = 1$,

i.e. $w=z+b$, the mapping reduces to a translation of the plane, the displacement vector being b.

7. Extended Complex Plane. Before proceeding to consider further examples we shall make some general remarks. In the theory of analytic functions, besides ordinary (*proper*) points of the complex z-plane we usually consider an additional **improper** point called the **point at infinity** (denoted as $z=\infty$). This point serves as the limit of any sequence of finite points $z_1, z_2, z_3, \ldots$, for which $|z_n| \underset{n\to\infty}{\longrightarrow} \infty$. The complex plane with the point at infinity (or, simply, **infinity**) added to it is referred to as the **extended complex plane.**

To visualize the extended complex plane, let us take a sphere (S) tangent to the complex plane and project all the points of the plane onto it by means of the rays passing through the point $P \in (S)$ opposite to the point of tangency O (see Fig. 5). This projection determines a one-to-one mapping of the complex plane on the sphere (S) with the point P deleted because no finite point of the plane corresponds to it. But if we consider an arbitrary sequence $A, B, C, \ldots$ of points in the plane which approach infinity, the corresponding points $A', B', C', \ldots$ of the sphere (S) tend to P (why?). Therefore it is natural to add the point at infinity to the complex plane and assume that, under this projection, it corresponds to the point P. The extended complex plane thus obtained is mapped in a one-to-one manner onto the whole sphere (S) *), and it is readily seen that both the mapping and its inverse are **continuous** (a mapping f is said to be continuous if $x \longrightarrow x_0$ always implies $f(x) \longrightarrow f(x_0)$). Note that the point at infinity, like the point zero, does not possess a uniquely specified value of the argument. A **neighbourhood of the point** $z=\infty$ is naturally understood as a part of the z-plane which is the exterior of a region bounded by a closed contour (think about it!).

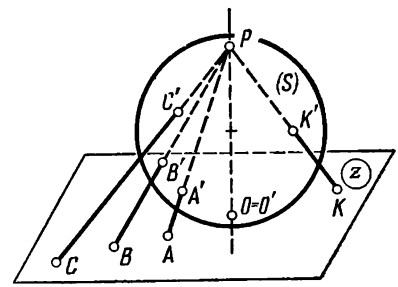


Fig. 5

If it is possible to establish a one-to-one correspondence between the elements of two sets in such a way that it is continuous in both directions we say that these sets are **homeomorphic** or **topologically equivalent**. There is a division of mathematics named **topology** (from the Greek τοπος place + λογος speech,

*) This construction is known as the **stereographic projection**, the sphere (S) being referred to as the **Riemann sphere**.— *Tr.*

thought) which studies the geometric properties of sets, manifolds and spaces preserved under the **homeomorphisms** (i.e. under one-to-one mappings continuous in both directions). From the point of view of topology, homeomorphic (topologically equivalent) objects are indistinguishable in the sense that all (topological) properties of one of them are common to the others. For instance, from the topological point of view, a circle, an ellipse and a polygon are equivalent but they are not equivalent to a line segment or a disk. Thus, the conclusion drawn in the foregoing paragraph can be formulated as follows: *the extended complex plane and the Riemann sphere are homeomorphic.*

We see that the symbol ∞ is used in two senses: as "improper real number ∞" (or $+\infty$) and "improper complex number ∞" which are not the same. Usually it is clear from the context which sense is meant but there are also some ambiguous situations where the corresponding stipulation is necessary.

8. Fractional Linear Mapping. The **fractional linear** mapping is defined by the formula

$$w=\frac{az+b}{cz+d} \tag{11}$$

where a, b, c and d are given complex numbers (it is also known as the **Möbius transformation**). Let us begin with the special case

$$w=\frac{k^2}{z}. \tag{12}$$

where $k>0$ is real. Passing to the modulus and argument we obtain

$$|w|=\frac{k^2}{|z|},\quad \operatorname{Arg} w=-\operatorname{Arg} z \tag{13}$$

For simplicity, let us first assume that the argument remains unchanged and only the modulus varies. This means that we take the auxiliary mapping $s=s(z)$ for which

$$|s|=\frac{k^2}{|z|},\quad \operatorname{Arg} s=\operatorname{Arg} z \tag{14}$$

Regarding the z-plane and the s-plane as coincident and taking into account the second equality (14) we can say that every ray issued from the origin is transformed into itself (why?). Let us investigate this transformation. It follows from the first equality (14) that for $|z|=k$ we have $|s|=k$ and for $|z|<k$ we obtain $|s|>k$, the quantities $|z|$ and $|s|$ being in inverse variation so that $|z|\to 0$ implies $|s|\to\infty$. For $|z|>k$ the situation is analogous. In Fig. 6 we see several points and their images, the corresponding points being supplied with the same indices. It should

be noted that $|s|=\frac{k^2}{|z|}$ implies $|z|=\frac{k^2}{|s|}$ and vice versa, which means that when a preimage goes into its image the image goes into the preimage. In other words, the mapping in question transforms the plane into itself and is the inverse of itself. Mappings possessing this property are called **involutory** (or, simply, **involutions**). Under the mapping in question all the points of the circle $|z|=k$ go into itself, i.e., remain fixed (this is **the invariant circle** of the mapping). The points of concentric circles of radii less than k are transformed into the points of the corresponding circles of radii greater than k and vice versa (see Fig. 6). The points 0 and ∞ go into each other under this mapping.

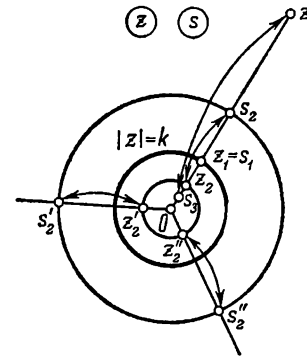


Fig. 6

Mapping (14) is termed an **inversion** *or* **reflection** *of the z-plane in the circle* $|z|=k$ (*or* **symmetry transformation with respect to the circle** $|z|=k$), and the points going into one another under an inversion are said to be **symmetric with respect to the circle**. This terminology is accounted for by the fact that in any sufficiently small neighbourhood of every point of the invariant circle this mapping coincides, to within infinitesimals of higher order, with an ordinary (mirror) reflection. (It can also be shown that if the point O in Fig. 5 is the origin and the sphere (S) is of radius $\frac{k}{2}$, mapping (14), when transferred to this sphere, induces the reflection of the sphere (S) through the plane parallel to the z-plane and containing the centre of the sphere.)

Let us proceed to prove the following property of inversion known as **circle-preserving property**: *under an inversion all the circles and straight lines are transformed into circles and straight lines*, and every circle (straight line) is transformed either into a circle or into a straight line. To prove this assertion we note that the equations of all the circles and straight lines in the xy-plane can be written in the form

$$\alpha(x^2+y^2)+\beta x+\gamma y+\delta=0 \qquad (15)$$

because for $\alpha\neq 0$ relation (15) is the equation of a circle (why?) and for $\alpha=0$ we obtain the equation of a straight line. Putting $z=x+iy$ and $s=p+iq$ we can write down the following formulas for mapping (14) (whose derivation is left to the reader):

$$x=\frac{k^2p}{p^2+q^2}, \qquad y=\frac{k^2q}{p^2+q^2}$$

Substituting these expressions into equation (15) we obtain the equation of a line which is the image of line (15), and after some simplifications we arrive at the relation (check it up!)

$$\delta(p^2+q^2)+\beta k^2 p+\gamma k^2 q+\alpha k^4=0$$

Since $k=\text{const}$, the latter equation is again of the form (15) but with other coefficients. Hence, we have obtained a circle or a straight line. (Let the reader prove that the image is a straight line only if the original circle or straight line passes through the origin.)

Mapping (13) is obtained from mapping (14) by the additional mapping $|w|=|s|$, $\operatorname{Arg} w=-\operatorname{Arg} s$, i.e. $w=s^*$. Hence, mapping (13) is a combination (superposition) of an inversion and reflection of the plane through the real axis. The function (12) being analytic everywhere except the point $z=0$, the mapping performed by this function is conformal throughout the plane with the point $z=0$ deleted, and thus the plane with deleted origin is conformally mapped onto itself. By the way, this can also be regarded as a conformal mapping of the extended complex plane onto itself. As for inversion, it is a conformal mapping of the second kind (see Sec. 5).

Now let us proceed to consider the general fractional linear mapping (11). Transforming the right-hand side according to the formula

$$w=\frac{a}{c}\,\frac{z+\frac{b}{a}}{z+\frac{d}{c}}=\frac{a}{c}\left(1+\frac{\frac{b}{a}-\frac{d}{c}}{z+\frac{d}{c}}\right)=\left(\frac{b}{c}-\frac{ad}{c^2}\right)\frac{1}{z+\frac{d}{c}}+\frac{a}{c}$$

we see that w can be obtained from z by applying, in succession, the mappings

$$r=z+\frac{d}{c},\quad t=\frac{1}{r}\quad\text{and}\quad w=\left(\frac{b}{c}-\frac{ad}{c^2}\right)t+\frac{a}{c}$$

If all the four complex planes z, s, t and w are considered to be coincident, the mapping in question is a combination of translation, inversion, reflection in a straight line and a linear mapping (see Sec. 6). All these mappings possessing the circle-preserving property, we conclude that the general Möbius mapping possesses it as well. We can also assert that *this mapping transforms points symmetric with respect to a circle (a straight line) into points symmetric with respect to the image of this circle (straight line).*

Expressing the variable z from (11) in terms of w we see that *the inverse of a fractional linear mapping is again a fractional linear mapping* although, in contrast to (12), it does not necessarily coincide with the original mapping in the general case. Similarly,

it is readily proved that a superposition (i.e. successive application) of fractional linear mappings results in a mapping of the same type.

The right-hand side of (11) involves four complex parameters but if they are multiplied by a factor the function (11) and the corresponding mapping remain unchanged; therefore here we only have three independent complex parameters ("complex degrees of freedom") or six real independent parameters ("real degrees of freedom") because one complex parameter involves two real parameters (on the notion of a degree of freedom see, for example, [107], Sec. X.2). Hence, in order to determine a linear fractional mapping uniquely we must set three conditions connecting the complex parameters. For instance, we can arbitrarily choose three values of z and set three (different!) values of w corresponding to them. Then it is readily seen that mapping (11) is determined uniquely.

Consider an example. Let a mapping of form (11) transform the points $z=0,\ 1,\ \infty$ into the corresponding points $w=-1,\ -i,\ 1$. Substituting these values into (11) we obtain the equalities

$$\frac{b}{d}=-1,\quad \frac{a+b}{c+d}=-i,\quad \frac{a}{c}=1$$

Expressing all parameters in terms of one of them and substituting the expressions thus obtained into (11) we finally receive the sought-for mapping (check it up!):

$$w=\frac{z-i}{z+i} \tag{16}$$

The circle-preserving property implies that this mapping transforms the "circle" passing through the three given points $z=0,\ 1,\ \infty$, that is the real axis *), into a circle which passes through the points $w=-1,\ -i,\ 1$, that is into **unit circle** $|w|=1$. Since a conformal mapping preserves orientation we conclude that the **upper half-plane** $\operatorname{Re} z>0$ is mapped onto the **unit disk** $|w|<1$. (Let the reader find out what is the image of the **lower half-plane** $\operatorname{Re} z<0$; see Fig. 7.)

Mapping (16) is not the only one transforming conformally the upper half-plane into the unit disk because we can choose any other three points of the unit circle (taken in the order they are met when the circle is traversed in the positive direction) as images of the same points $z=0,\ 1,\ \infty$. Consequently, the totality of these mappings possesses three real degrees of freedom ("one and a half complex degrees of freedom"). It can be shown that the general

*) A straight line is naturally regarded as a limiting case of a circle corresponding to infinite radius.— *Tr.*

form of such a mapping is $w = e^{i\alpha}\frac{z-p}{z-p^*}$ where α is an arbitrary real number and p is any complex number such that $\operatorname{Im} p > 0$.

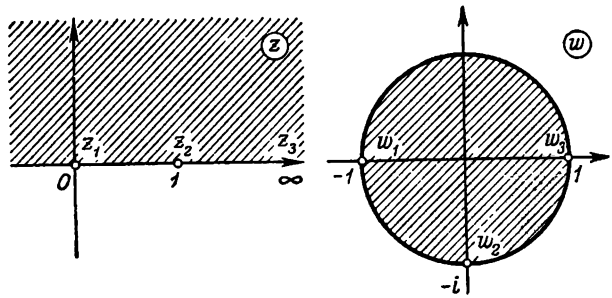


Fig. 7

9. The Mapping $w=z^k$. Let us consider the mapping

$$w = z^k \tag{17}$$

where k is a given real number. We shall begin with the special case

$$w = z^2 \tag{18}$$

Taking the modulus and the argument we can write

$$|w| = |z|^2, \quad \operatorname{Arg} w = 2 \operatorname{Arg} z \tag{19}$$

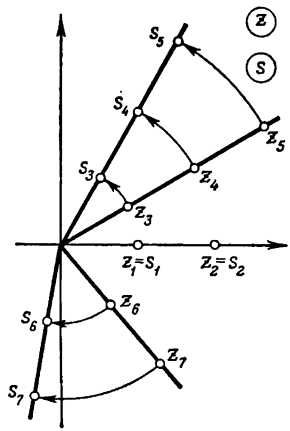


Fig. 8

Now, for simplicity, imagine that the modulus remains unchanged, that is consider the auxiliary mapping $s = s(z)$ for which

$$|s| = |z|, \quad \operatorname{Arg} s = 2 \operatorname{Arg} z \tag{20}$$

Under this mapping every point is turned about the origin (why?), and the points having the same argument before the mapping, i.e. lying on a common ray starting from the origin, go into points which again belong to a common ray, the argument becoming twice as large (see Fig. 8). Consequently, every ray of that type, when mapped, is turned about its vertex as a rigid rod. If we now consider all such rays lying in a fixed angle we see that the situation is as if a "fan" consisting of the rods and occupying this initial angle with vertex at the origin is uniformly spread in such a way that this angle is doubled. For this fan not

to overlap itself in this "spreading", it is necessary that the initial angle be not greater than 180°, i.e. the preimage must lie within a straight angle.

Now let us come back to mapping (19). Comparing it with mapping (20) we see that here, in addition to (19), the points for which $|z|<1$ move towards the origin because for them $|w|=|z|^2<|s|=|z|$, whereas the points for which $|z|>1$ move away from the origin. Thus, when mapped, the points slide along the rotating rays. Hence, the function (18) conformally maps every angle of magnitude $\alpha<\pi$ with vertex at the origin onto another angle of magnitude 2α. The mapping is conformal everywhere except the origin where $\frac{dw}{dz}=2z\Big|_{z=0}=0$ (cf. Sec. 5). Under the mapping, every angle with vertex at the point $z=0$ becomes twice as large; this again indicates that the mapping is not conformal at the origin.

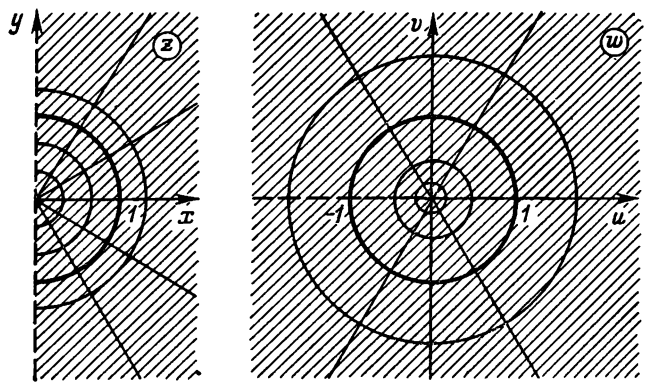


Fig. 9

If we pass to the case $\alpha=\pi$ then what has been said in the preceding paragraph must be formulated in a more precise manner. Namely, the original domain should be regarded as *being open*, i.e. its boundary must not be included into it. Then the image (the domain obtained by the mapping) is also open, and thus we arrive at the mapping shown in Fig. 9 where the image and preimage (we mean the corresponding domains) are shaded. The boundaries which are not included into the domains are shown in dotted lines, and the positions of some other lines before and after the mapping are also depicted there. We thus conclude that function (18) performs a conformal and univalent mapping of the open *right half-plane* of the variable z (i.e. the half-plane $\operatorname{Re} z>0$) onto the w-plane with the ray $v=0$, $-\infty<u\leqslant 0$ deleted or, as

we say, on the w-plane with a **cut** made along that ray. If we considered the closed half-plane, the boundary would overlap itself after the mapping (why?), and hence the mapping would no longer be univalent (see Sec. 4).

The angle of π radians in the z-plane should not necessarily be in the position shown in Fig. 9. For instance, let the reader show that the same function (18) determines a conformal mapping of the open upper (lower) half-plane of the variable z onto the w-plane with the cut along the positive half of the real axis.

If function (18) is regarded as defined throughout the z-plane, the w-plane is covered twice by the points $w=z^2$ (why?). Consequently, this mapping is single-valued but two-sheeted (2-valent; see Sec. 4). It is seen from Fig. 9 that when the point z traverses one circuit about the point $z=0$ the corresponding point w describes two circuits round the point $w=0$.

In the general case $k>1$ the mapping determined by function (17) has a similar structure. This mapping transforms conformally every open angle of magnitude $\frac{2\pi}{k}$ with vertex at the origin lying in the z-plane onto the whole w-plane with cut along a ray drawn from the point $w=0$. It should be noted that if k is not an integral number the situation is more complicated because then function (17) is no longer one-valued (why?). In this case we can take a concrete value of z and choose one of the values of z^k corresponding to it (for instance, we can put $w=1$ for $z=1$) and obtain the values of w for all the other points z by extending the chosen value in a continuous manner (cf. Sec. 10). If k is an integer and (17) is regarded as a function defined throughout the z-plane, the corresponding mapping is k-sheeted (k-valent).

If $0<k<1$, function (17) performs a conformal mapping of the z-plane with cut along a ray starting from the origin onto the corresponding open angular region with angle of extension $2\pi k$ and vertex at the origin (in the w-plane). Let the reader consider the case $k<0$.

10. Multiple-Valued Functions and Branch Points. Let us consider in more detail function (17) for the case $k=\frac{1}{2}$, i.e. $w=\sqrt{z}$. This function is two-valued. It would be incorrect to write, by analogy with real radicals, the two values of $w=\sqrt{z}$ in the form $w_1=+\sqrt{z}$ and $w_2=-\sqrt{z}$ because the notion of an "arithmetic" root of a complex number is senseless. To distinguish between the values of $\sqrt{z}$ we must indicate which *branch* of this two-valued function is considered. Similarly, the relation $\sqrt{-z}=i\sqrt{z}$ can only be meaningful if we indicate which values of the root are taken, etc.

We see that $w=\sqrt{z}$ takes on the two values $w=1$ and $w=-1$ for $z=1$. Choose one of them, for instance, $w=1$. Now let us vary z and continuously extend the corresponding values of $\sqrt{z}$. For example, let the point z traverse the unit circle in the positive direction and pass, in succession, through the points $z_0, z_1, z_2, \ldots$ shown in Fig. 10. Since $\operatorname{Arg} w=\frac{1}{2}\operatorname{Arg} z$, the point w moves twice as slow; when the point z completes a circuit about the origin and arrives at the point $z_4=z_0=1$, the point w traverses only a half-circuit and comes to the position $w_4=-1$. We see that if we start

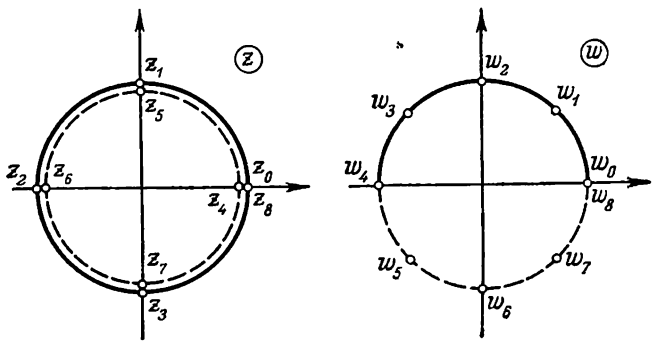


Fig. 10

from the chosen value $\sqrt{z}|_{z=1}=1$ and extend the values of $w=\sqrt{z}$ in a continuous manner we necessarily arrive at the other value $\sqrt{z}|_{z=1}=-1$. If now the point z describes one more closed circuit round the origin (it is shown in dotted line in Fig. 10) the corresponding point w traverses another half-circuit and returns to the original value $w_8=w_0=1$.

It is clear that the same situation occurs when the point z makes circuits round the origin along any closed contour other than the unit circle.

The above example demonstrates a typical situation and indicates an essential difference between multiple-valued functions of a real and a complex variable. For a many-valued function of a real variable we can introduce, in a natural way, its single-valued branches which are different functions (e.g. see [107], Sec. I.20). For instance, the expression $\sqrt{x}$ is always understood as the "arithmetic" branch of the two-valued function $\pm\sqrt{x}$. In contrast to it, the values of a multiple-valued function of a complex variable are, as a rule, closely interrelated and continuously go into one another when the point representing the independent

variable describes closed contours enveloping certain points which are called **branch points** of the given function. As has been shown, for the two-valued function $w=\sqrt{z}$ the origin is its branch point. By the way, the point $z=\infty$ is also considered to be a branch point of this function. Indeed, "to traverse a circuit about infinity" means to describe a circle of large radius with centre at a finite point (see Fig. 5), the values of $\sqrt{z}$ continuously changing into one another.

Hence, if we want to specify continuous single-valued branches of a multiple-valued function $f(z)$ and consider them separately we must prevent the point z from making circuits about the branch points of the function. For this purpose we usually make cuts joining the branch points in the z-plane (cf. Sec. 9) in such a way that if the path of the point z does not intersect the cuts, the values of $f(z)$ corresponding to different branches cannot go into one another. As an example, let us take Fig. 9 and change the roles of the planes shown there, that is denote the z-plane as w-plane and vice versa (the same can be achieved if we interchange the independent and dependent variables without changing the notation). Then we have a cut on the z-plane (which now, under the above convention, is shown in the right Fig. 9) connecting the branch points $z=0$ and $z=\infty$. In the plane with such a cut we can define a single-valued branch of the function $w=\sqrt{z}$ which assumes the value $w=1$ for $z=1$. This branch determines a conformal mapping of the plane with the cut onto the right half-plane w (Fig. 9). The other branch for which $w|_{z=1}=-1$ determines a mapping of the z-plane with the same cut onto the left half-plane w. For each branch the cut is a line of discontinuity because on the edges (lips) of the slit each branch takes on the corresponding limiting values and thus has a finite jump (jump discontinuity). Fig. 9 shows that the branch depicted there takes on at the point $z=-1$ the value $w=-i$ on the lower edge of the cut and the value $w=i$ on the upper edge.

Similarly, the function $w=\sqrt[n]{z}$ where $n=2, 3, 4, \ldots$, is n-valued and has the branch points $z=0$ and $z=\infty$. The values of the function corresponding to different branches continuously change into one another when the point z describes closed contours around them. After the point has traversed n circuits about its branch point all the values of the function return to their original values. A point of this type is said to be a **branch point of order** n *). The function $w=z^{m/n}$ where $n=2, 3, 4, \ldots$, and m is an integer without common factors with n also has two branch points $z=0$

*) In this case the point is also called an **algebraic branch point** provided that the function has a limit (finite or infinite) at it.—*Tr.*

and $z=\infty$ each of order n. Consequently, the greater the denominator, the greater the number of different values of the function, i.e. the higher the order of the branch point. The function $w=z^k=e^{k\ln z}$ where k is an irrational number is infinite-valued, and when the point z passes around its branch points $z=0$ and $z=\infty$ the values of the function go into one another but never come back to the original value. These branch points are called **logarithmic** or are said to be **branch points of infinite order.**

The existence of branch points is a characteristic feature of multiple-valued analytic functions. If we consider the function $w=\sqrt{z^2}=\pm z$ we see that it is two-valued but its branches $w=z$ and $w=-z$ do not change into one another when they are continuously extended, and thus this function cannot be considered a "single" multiple-valued analytic function; from the point of view of the theory of analytic functions it is an artificial aggregate made of two different functions $w=z$ and $w=-z$. Similarly, if the function $w=\sqrt{z}$ is considered in a small neighbourhood of a point $z_0\neq 0,\ \infty$ its two branches do not continuously turn into one another and thus do not form a "single" analytic function. But it can be proved that *if a multiple-valued analytic function is defined in a simply connected domain* (in particular, in the whole complex plane), *and its values can be made to change into one another when they are continuously extended along closed contours, the function must have at least one branch point in this domain.* (Let the reader examine the role of the condition that the domain is simply connected.)

For functions determined by simple formulas the branch points can usually be found by considering the values of the argument for which the expressions under the radical signs vanish or approach infinity (this makes it possible to find the branch points of finite orders) or the expressions under the sign of logarithm turn into zero or tend to infinity (logarithmic branch points). For instance, consider the function

$$w=\sqrt{z^2-4} \tag{21}$$

This is a two-valued function with branch points found from the equation $z^2-4=0$ which yields $z_{1,2}=\pm 2$. The expression (21) can be written in the form $w=\sqrt{z-2}\sqrt{z+2}$, and hence we see that if the point z describes a small circle about the point $z_1=2$, the point in the z-plane representing the number $z-2$ traverses a circuit about the origin, and the first factor passes to its new value whereas the second remains the same. Hence, the whole function (21) changes its value. After a second circuit both factors together with the function return to the original values. Thus, the point z_1 and, similarly, the point z_2 are branch points of order

two for the function (21). The point $z=\infty$ is not a branch point of function (21) although the radicand approaches infinity at it. Virtually, the function in question can be represented as

$$w=\sqrt{z^2\left(1-\frac{4}{z^2}\right)}=\sqrt{z^2}\sqrt{1-\frac{4}{z^2}}=(\pm z)\sqrt{1-\frac{4}{z^2}}$$

If z describes a circle of large radius, both factors (and the whole function together with them) return to their original values (why?). Consequently, in the neighbourhood of the point at infinity there are two different branches which are not connected with each other.

Let the reader show that the function $\sqrt{z}+\sqrt[3]{z-1}$ is 6-valued and has the branch points $z=0$ (of the second order), $z=1$ (of the third order) and $z=\infty$ (of the sixth order).

11. The Mapping $w=\frac{c}{2}\left(z+\frac{1}{z}\right)$. The mapping

$$w=\frac{c}{2}\left(z+\frac{1}{z}\right)\quad (c>0) \tag{22}$$

is widely used in hydrodynamics and aerodynamics. Function (22) is called the **Zhukovsky function** after N. E. Zhukovsky who was the first to apply the methods of the theory of analytic functions to a wide class of problems encountered in these fields of science.

Let us introduce, in the plane $z=x+iy$, polar coordinates ρ, φ connected with x and y by the formulas $x=\rho\cos\varphi$, $y=\rho\sin\varphi$ which can be written in the form $z=\rho e^{i\varphi}$. Then, by (22), we have

$$w=\frac{c}{2}\left(\rho e^{i\varphi}+\frac{1}{\rho}e^{-i\varphi}\right)=\frac{c}{2}\left(\rho+\frac{1}{\rho}\right)\cos\varphi+i\,\frac{c}{2}\left(\rho-\frac{1}{\rho}\right)\sin\varphi=$$
$$=u+iv$$

and therefore

$$u=\frac{c}{2}\left(\rho+\frac{1}{\rho}\right)\cos\varphi,\quad v=\frac{c}{2}\left(\rho-\frac{1}{\rho}\right)\sin\varphi \tag{23}$$

If $\rho=\text{const}>1$ and φ varies from 0 to 2π, formulas (23) are parametric equations of an ellipse in the u v-plane (i. e. in the w-plane) with semi-axes

$$a=\frac{c}{2}\left(\rho+\frac{1}{\rho}\right)\quad\text{and}\quad b=\frac{c}{2}\left(\rho-\frac{1}{\rho}\right)$$

(e. g. see [107], equations (II.26)). Thus, the circle $\rho=\text{const}>1$ in the z-plane is mapped by the function (22) in a one-to-one manner onto the ellipse (23) lying in the w-plane (see Fig. 11). According to the well-known formula of analytic geometry (e. g. see [107], Sec. II.10), the half-distance between its focuses is

equal to

$$\sqrt{a^2-b^2}=\sqrt{\left[\frac{c}{2}\left(\rho+\frac{1}{\rho}\right)\right]^2-\left[\frac{c}{2}\left(\rho-\frac{1}{\rho}\right)\right]^2}=$$
$$=\frac{c}{2}\sqrt{\left(\rho^2+2+\frac{1}{\rho^2}\right)-\left(\rho^2-2+\frac{1}{\rho^2}\right)}=c$$

Consequently, the parameter c entering into expression (22) has been given a geometric interpretation.

Now, taking various circles $\rho=\text{const}>1$ we obtain in the w-plane the corresponding family of ellipses (Fig. 11), and since c is independent of ρ, the ellipses have the same focuses. For $\rho\to\infty$ we have $a\to\infty$ and $b\to\infty$ (and $a-b\to 0$), and for $\rho\to 1+0$ we have $a\to c+0$ and $b\to+0$. We see (Fig. 11) that function

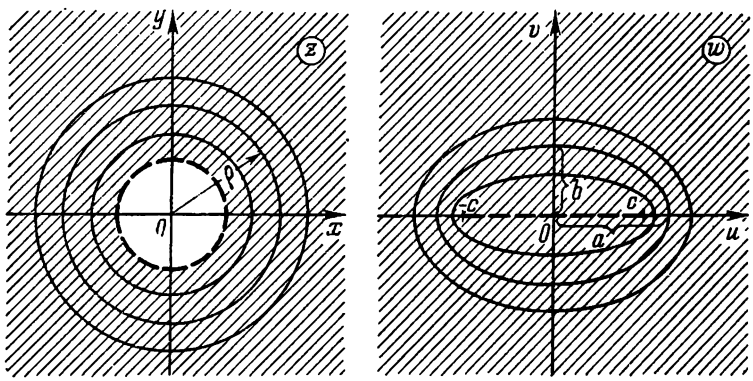


Fig. 11

(22) performs a conformal mapping of the exterior of unit circle lying in the z-plane onto the w-plane with cut along the line segment connecting the points $-c$ and c of the real u-axis. The unit circle itself undergoes a 2-valent mapping onto the cut; the mapping is conformal everywhere except the points $z=\pm 1$ at which $\frac{dw}{dz}=\frac{c}{2}\left(1-\frac{1}{z^2}\right)=0$ (this is clearly seen in Fig. 11). Let the reader show that the open unit circle with centre at the origin in the z-plane is mapped by function (22) onto the w-plane with the same cut (the point $z=0$ going into $w=\infty$). We now conclude that function (22) regarded as defined on the entire z-plane determines a 2-sheeted mapping of the z-plane on the w-plane, the inverse mapping being two-valued and having the two branch points $w=\pm c$ of the second order.

Combining simple conformal mappings we can considerably extend the class of domains which can be mapped onto one another. As

an example, consider the problem of constructing a mapping transforming the domain depicted in Fig. 12 into the half-plane $\operatorname{Re} w > 0$. Taking the inverse of mapping (22) for $c = 1$ and changing the notation we see that the function

$$p = z + \sqrt{z^2 - 1}$$

maps the given region onto the domain shown in Fig. 13 (here we have taken the single-valued branch of the square root which is positive for real $z > 1$). It is readily proved that the linear fractional mapping

$$q = \frac{p - i}{p + i}$$

transforms the domain shown in Fig. 13 into the angle $\operatorname{Re} q > 0$, $\operatorname{Im} q < 0$ in the complex q-plane. Now we take the mapping $w = iq^2$ and combining the formulas thus obtained arrive at the function

$$w = i\left(\frac{z + \sqrt{z^2 - 1} - i}{z + \sqrt{z^2 - 1} + i}\right)^2 \tag{24}$$

which determines the required mapping.

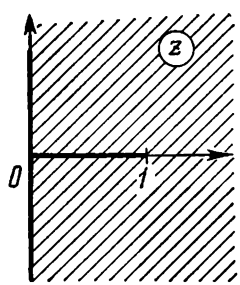


Fig. 12

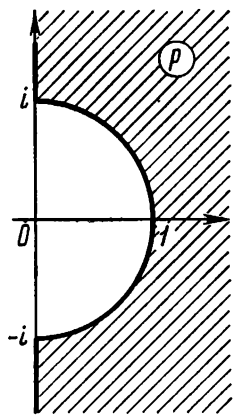


Fig. 13

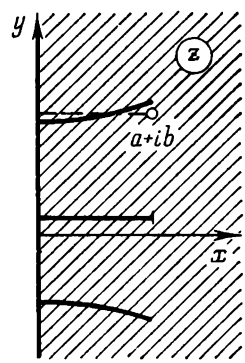


Fig. 14

The result we have obtained can be used for determining a mapping which approximately transforms the domain shown in Fig. 14 into a half-plane. The domain to be mapped is the half-plane $\operatorname{Re} z > 0$ with several cuts drawn from the y-axis which, on the one hand, are not placed too close to one another and, on the other hand, do not considerably deviate from the horizontal line segments $y = \text{const}$ shown in Fig. 14. To construct the mapping, we approximately replace one of the cuts by the line segment $y = b$, $0 < x < a$

(see Fig. 14) and perform the mapping $z_1 = \Phi\left(\frac{1}{a}(z - ib)\right)$ where Φ is the function defined by (24). Then the given domain goes into the half-plane $\operatorname{Re} z_1 > 0$ with cuts. These cuts are of the same type as before but their number is reduced by unity. Then we again approximately replace one of the new cuts by the line segment $y_1 = b_1$, $0 < x_1 < a_1$ and apply the mapping $z_2 = \Phi\left(\frac{1}{a_1}(z_1 - ib_1)\right)$, etc. until all the cuts are "straightened".

12. Exponential Mapping and Some Mappings Related to It. We shall consider the function

$$w = e^z$$

Putting $z = x + iy$ we write

$$w = e^x e^{iy}, \text{ i. e. } |w| = e^x, \operatorname{Arg} w = y + 2k\pi \qquad (25)$$

If $y = \text{const}$ and $-\infty < x < \infty$ we have $\operatorname{Arg} w = \text{const}$ and $0 < |w| < \infty$, and therefore the straight line $y = \text{const}$ is mapped in a one-to-one manner on a ray drawn from the origin in the w-plane (Fig. 15). Considering all such straight lines for different $y = \text{const}$ we obtain, as the collection of their images, a pencil

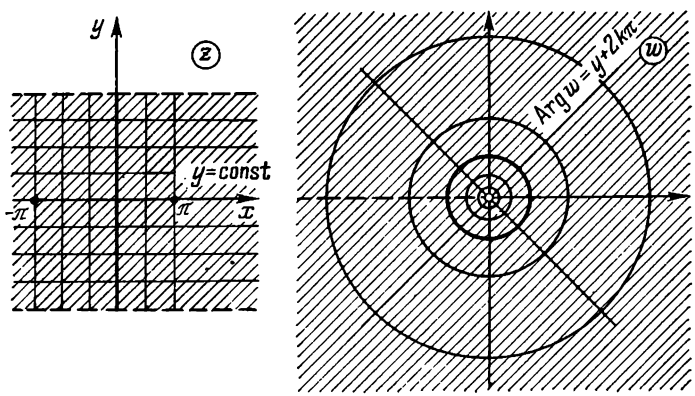


Fig. 15

of rays. For these rays to cover continuously the whole w-plane it is necessary that the polar angle by which every ray is specified range within an interval of length 2π; for instance, let $-\pi < y < \pi$, as is shown in Fig. 15. Consequently, function (25) determines a conformal mapping of the open strip of width 2π (Fig. 15), which is parallel to the real axis of the z-plane, on the w-plane with cut along a ray passing through the origin (in Fig. 15 this ray coincides with the negative real half-axis).

Under this mapping, the straight lines $x = \text{const}$ go into the circles $|w| = e^x = \text{const}$, that is the Cartesian coordinate lines of the z-plane are transformed into the polar coordinate lines in the w-plane. For $x \to -\infty$ we have $w \to 0$, and thus the left portion of the strip "lying at infinity" goes into the vertex of the pencil of rays at the origin in the w-plane while the right portion "at infinity" is transformed into the point at infinity of the w-plane.

If mapping (25) is considered on the entire z-plane, every horizontal strip of width 2π in the z-plane is mapped onto the whole w-plane, and hence the w-plane is covered by the points $w = e^z$ infinitely many times when z runs throughout the z-plane. Thus, mapping (25) is infinitely-sheeted. The corresponding inverse mapping $z = \operatorname{Ln} w$ is infinite-valued and has the branch points $w = 0$ and $w = \infty$ of infinite order: if the point w describes circuits about the origin of the w-plane in the positive direction, the function $\operatorname{Ln} w$ gains the increment $2\pi i$ after every circuit and never returns to the original value. If we draw a cut in the w-plane connecting the branch points, we can separate the single-valued branches of the logarithmic function. For example, in Fig. 15 we see a conformal mapping of the w-plane with cut along the negative real axis onto the strip $-\infty < x < \infty$, $-\pi < y < \pi$ in the z-plane. The corresponding branch is termed the **principal value of the logarithmic function:** $z = \ln w = \ln |w| + i \arg w$. (Remember that the symbol $\arg w$ designates the principal value of the argument: $-\pi < \arg w \leqslant \pi$.)

There are some other mappings that are closely related to the exponential mapping. For instance, the mapping

$$w = \cosh z = \frac{e^z + e^{-z}}{2}$$

can be obtained as a combination of two mappings, namely

$$s = e^z \quad \text{and} \quad w = \frac{1}{2}\left(s + \frac{1}{s}\right)$$

The first of them (see Fig. 15) maps the semi-strip $0 < x < \infty$, $-\pi < y < \pi$ onto the domain $|s| > 1$ with cut along the line segment $s'' = 0$, $-\infty < s' < -1$ where $s' = \operatorname{Re} s$ and $s'' = \operatorname{Im} s$ (i. e. $s = s' + is''$). The second mapping transforms this domain into the w-plane with cut along the segment $[-1, 1]$ of the real axis (see Fig. 11). Then the segment $s'' = 0$, $-\infty < s' < -1$ goes into the segment $v = 0$, $-\infty < u < -1$ ($u = \operatorname{Re} w$, $v = \operatorname{Im} w$, $w = u + iv$) and hence the resultant transformation is a conformal mapping shown in Fig. 16 where the points of the boundaries of the domains in question which correspond to each other are marked with the same indices. Let the reader study the mapping $w = \cosh z$ considered on the entire w-plane.

The mappings

$$w=\cos z=\cosh iz,\ w=\sin z=\cos\left(\frac{\pi}{2}-z\right)\ \text{and}$$

$$w=\sinh z-\frac{1}{i}\sin iz=-i\cosh\left(z+\frac{\pi i}{2}\right)$$

possess similar properties.

In this section we limit ourselves to the examples given above. Another important class of conformal mappings will be studied in Sec. 3.13. For more detail concerning these and other elementary conformal mappings we refer the reader to [41], [42], [81],

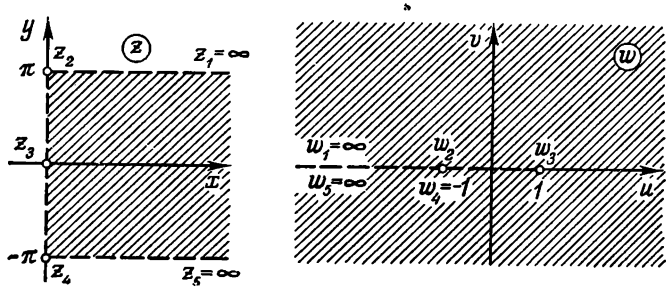


Fig. 16

[88], [93], [105], [120], [123], [126]. There are some special monographs, for instance [66], where the reader can find catalogues of conformal mappings.

13. Riemann Surface. The notion of a *Riemann surface* proves to be useful in investigating multiple-valued analytic functions. Let us turn back to the two-valued function $w=\sqrt{z}$ (Sec. 10) and imagine that we have two "replicas" of the z-plane, with the cuts along the half-axis $y=0$, $-\infty<x<0$, which are made to coincide. Suppose the point z can arbitrarily move throughout the two replicas of the z-plane (called **sheets**) on condition that when its path intersects the (coinciding) cuts it *passes from one sheet to the other*. In other words, we think of these sheets as being "pasted together" in such a way that the upper edge of the cut in the first sheet is connected to the lower edge of the cut in the second sheet and the lower edge of the cut in the first sheet joins the upper edge of the cut in the second sheet. This configuration is conventionally shown in Fig. 17 although it is physically unrealizable (because if we try to construct these sheets of paper there appear lines of self-intersection seen in Fig. 17). But in mathematics we sometimes consider surfaces or, generally, manifolds which are not necessarily realizable as ordi-

nary surfaces embedded in the real three-dimensional physical space. Therefore, nothing prevents us from thinking that after the sheets have been "pasted together" as described above the upper edge of the slit in the first plane is only connected with the lower edge of the slit of the second plane but not of the first one while the upper edge of the second plane adjoins the lower edge of the cut of the first plane but not of the second. Thus we imagine that there are no lines of self-intersection. The surface (S) thus obtained is called the **Riemann surface of the function** $w=\sqrt{z}$.

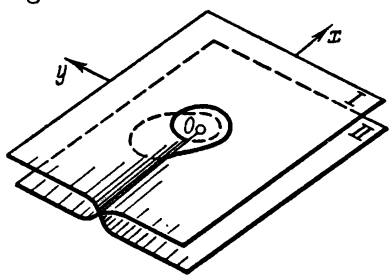


Fig. 17

Now consider a function, defined on that Riemann surface, which is equal to one of the branches of the function $\sqrt{z}$ (see Sec. 10) on one of the sheets of the surface and to the other branch of $\sqrt{z}$ on the other sheet. This function is one-valued and continuous on (S) because, as was shown in Sec. 10, these branches continuously change into each other when the point z passes across the cut; therefore the continuity is due to the fact that when the path of the point z intersects the ray along which the sheets are pasted together the point itself passes from one sheet to the other (think about it!). It is clearly seen in Fig. 17 that if the point z moves around the origin O the function $\sqrt{z}$ arrives at its original value only after two circuits are described. Thus, the function $w=\sqrt{z}$ (and therefore the function $z=w^2$ as well) determines a one-to-one correspondence, continuous in both directions, between the points z of the Riemann surface and the points of the complex w-plane.

We can similarly construct the Riemann surface for the three-valued function $\sqrt[3]{z}$ on which it becomes one-to-one and continuous. For this purpose we take three replicas of the z-plane (with the same cuts as above) and paste them together, in a cyclic order, in such a way that the upper edge of the slit on the first sheet is connected with the lower edge of the slit on the second sheet, the upper edge on the second sheet is connected to the lower edge on the third sheet and, finally, the upper edge on the third sheet adjoins the lower edge on the first sheet.

The Riemann surface of an infinite-valued function involves an infinitude of replicas of the z-plane which are pasted to one another in a manner that has been described above and form an infinitely-sheeted manifold which can be thought of as a sequence of sheets infinite in both directions.

The construction of that type can be made for any multiple-valued function $f(z)$. To obtain a Riemann surface corresponding to the given function we must take the number of replicas of the z-plane equal to the number of values the function assumes for each z (this number is the same for all z's except the branch points). Then we must make cuts to construct one-valued branches of the function and paste together these sheets along the cuts in the order in which the branches change into one another when the path of the point z intersects the cuts. The function $f(z)$ considered on its Riemann surface is one-valued and has no jumps when the point z crosses the cuts which are now pasted together in the appropriate order while its one-valued branches (considered on the "ordinary" z-plane) have jumps.

If a function $w=g(z)$ is single-valued and analytic in the entire z-plane, except possibly at separate points of discontinuity, but is multi-valent, it determines a one-to-one mapping of the z-plane onto the Riemann surface of the function $g^-(w)$ (the inverse function of $g(z)$). The mapping is conformal everywhere except the points of discontinuity of the function $g(z)$ and the points of discontinuity and branch points of the function $g^-(w)$. If a function $g(z)$ is both many-valued and multi-valent it specifies a conformal mapping of its Riemann surface on the Riemann surface corresponding to its inverse function.

It was indicated in Sec. 7 that the complex plane with the point at infinity added to it (the extended complex plane) is topologically equivalent to the Riemann sphere. It can be proved that, analogously, the Riemann surface, of a finite-valued function, extended in the same manner is topologically equivalent to a sphere, or to a sphere with a handle (which, in its turn, is equivalent to a torus) or to a sphere with two handles, etc. The Riemann surfaces can be classified according to the number of these handles. A Riemann surface corresponding to an infinite-valued function may be of a more complicated structure but in some special cases it can be simple. For instance, we see from Sec. 12 that the Riemann surface of the function $\operatorname{Ln} z$ is topologically equivalent to the plane.

14. Application to the Theory of Plane Fields. Let us consider a region (G) in the xy-plane in which a vector field $\mathbf{A}$ is defined. Suppose that the field $\mathbf{A}$ has neither vortices (curls) nor sources of vector lines, that is

$$\operatorname{rot} \mathbf{A}=0, \quad \operatorname{div} \mathbf{A}=0 \tag{26}$$

A field of this kind can be interpreted physically in various ways but for our further aims it is convenient to regard it as a velocity field of a plane-parallel flow of an incompressible **perfect**

(**ideal**) liquid (i. e. without viscosity which always generates vortices).

As was shown in Sec. I.2.3, a field satisfying conditions (26) is the gradient field of a harmonic function $\varphi(x, y)$. According to Sec. 3, it is possible to construct a function $\psi(x, y)$ which is the harmonic conjugate of $\varphi(x, y)$. The function $\psi(x, y)$ is then called a **stream function (flow function)**, and the expression $w = \varphi + i\psi$, which is an analytic function of the complex variable $z = x + iy$, is referred to as the **complex potential** of the field **A**.

Let us discuss the meaning of the derivative $\frac{dw}{dz}$. According to Sec. 2, $\frac{dw}{dz}$ can be computed by putting $dz = dx$, and hence

$$\frac{dw}{dz} = \frac{\partial(\varphi + i\psi)}{\partial x} = \frac{\partial \varphi}{\partial x} - i\frac{\partial \varphi}{\partial y}$$

because, by the Cauchy-Riemann equations (6), we have $\frac{\partial \psi}{\partial x} = -\frac{\partial \varphi}{\partial y}$. But $\mathbf{A} = \operatorname{grad} \varphi$, i. e. $A_x = \frac{\partial \varphi}{\partial x}$ and $A_y = \frac{\partial \varphi}{\partial y}$, and consequently

$$\frac{dw}{dz} = A_x - iA_y$$

Therefore we have $\left(\frac{dw}{dz}\right)^* = A_x + iA_y$ and thus the complex number $\left(\frac{dw}{dz}\right)^*$ $\left(\text{but not } \frac{dw}{dz}!\right)$ represents the vector **A** of the field at each point z (why?).

In particular, it follows that even if the analytic function w thus constructed is multiple-valued, its derivative $\frac{dw}{dz}$ must be one-valued, and hence any two single-valued branches of the function $w(z)$ only differ by a constant summand from each other. Conversely, every analytic function $w(z) = \varphi + i\psi$ possessing a single-valued derivative $\frac{dw}{dz}$ serves as a complex potential of a plane vector field **A** satisfying condition (26). Thus, the investigation of plane fields satisfying conditions (26) reduces to the theory of analytic functions, which simplifies the problem.

Let us consider a **level line** of the stream function whose equation is $\psi(x, y) = \text{const}$. By Sec. 4, this line is orthogonal at every point M to the line $\varphi = \text{const}$ passing through the point, and the vector $\mathbf{A} = \operatorname{grad} \varphi$ is also orthogonal to that line $\varphi = \text{const}$ (e. g. see [107], Sec. XII.4); thus we conclude that the vector **A** is tangent to the line $\psi = \text{const}$ (Fig. 18) at the point M. But, since M is an arbitrary point, the family $\psi = \text{const}$ thus consists of vector lines of the field **A**. In particular, for a velocity field these lines are **stream lines** (the trajectories of the particles), which accounts for the term "stream function".

Consider another consequence of formula (7). Let us join two points M_1 and M_2 by an arbitrary curve (l) (see Fig. 18). Then we have

$$\psi_1-\psi_2=\int\limits_{(l)} d\psi=\int\limits_{(l)}\frac{\partial\psi}{\partial l}\,dl=\int\limits_{(l)}\frac{\partial\varphi}{\partial n}\,dl=\int\limits_{(l)}(\operatorname{grad}\varphi)_n\,dl=\int\limits_{(l)}A_n\,dl$$

which means that *the increment of the stream function along any arc (l) is equal to the flux of the field across (l) (from left to right relative to the direction in which the curve (l) is described).*

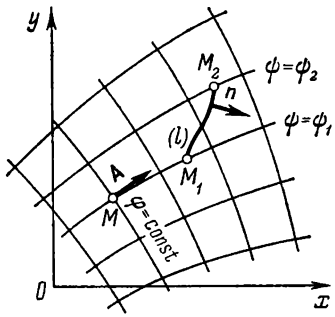


Fig. 18

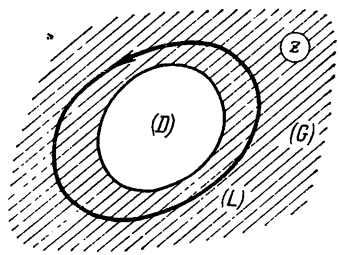


Fig. 19

Now we can give an interpretation to the situation when the function $w(z)$ is many-valued (this can only be the case if the domain (G) is multiply connected (cf. Sec. I.2.2). Let the point z traverse, in the positive direction, a closed contour (L) enclosing a "lacuna" (D) in the domain (G) (Fig. 19). In particular, this "lacuna" may consist of a single point. Then, according to Sec. I.2.2, the increment $\Delta_L\varphi=\Gamma_D$ of the function φ is equal to the circulation of the field $\mathbf{A}$ over the contour (L). But, by the preceding paragraph, the increment $\Delta_L\psi=Q_D$ of the function ψ is equal to the flux of the field $\mathbf{A}$ through (L) in the direction from the interior of (L) to the exterior, and consequently the increment of w is equal to $\Delta_L w=\Delta_L\varphi+i\Delta_L\psi=\Gamma_D+iQ_D$. Thus, *if the function w is multiple-valued, there are sources of circulation or sources of vector lines of the field $\mathbf{A}$ in the lacunas of the domain (G).* (It should be noted that there are no such sources in the interior of the domain (G), which is implied by condition (26).)

Now suppose that the domain (G) is finitely connected, for instance, k-connected $(2\leqslant k<\infty)$. Then there are $k-1$ lacunas (D_1), (D_2), $\ldots$, (D_{k-1}) in (G). Choose a closed contour (S) in (G) which envelopes all the lacunas and describe it in the negative direction. Then the quantity Γ_∞ entering into the increment $\Delta_S w=\Gamma_\infty+iQ_\infty$ is naturally interpreted as the circulation of the field "generated

at infinity", and the quantity Q_∞ as the strength of sources of vector lines "lying at infinity". Now it can easily be shown (how?) that

$$\Gamma_{D_1}+\Gamma_{D_2}+\ldots+\Gamma_{D_{k-1}}+\Gamma_\infty=0$$

and

$$Q_{D_1}+Q_{D_2}+\ldots+Q_{D_{k-1}}+Q_\infty=0$$

If the domain (G) is finite, the role of infinity is played by the outer connected component of the boundary of (G). If the domain (G) is not finite and some of the components of its boundary recede to infinity (for instance, such is the boundary of an infinite strip from which a disk is cut out: it has two infinite and one finitely connected components), then, when considering the quantities Γ_∞ and Q_∞, we regard these components as being "attached" to the point at infinity.

In conclusion we note that when we say that the function $w(z)$ is analytic we mean the interior of the domain (G) because the analyticity may be violated at some points of the boundary of (G).

15. Examples. Consider the complex potential

$$w=\frac{\Gamma}{2\pi i}\operatorname{Ln} z=\frac{\Gamma}{2\pi}\operatorname{Arg} z-i\frac{\Gamma}{2\pi}\ln|z| \tag{27}$$

where Γ is a real number. Equating the expression $\psi=\operatorname{Im} w$ to an arbitrary constant we obtain the equation $\ln|z|=\text{const}$ of the stream lines which thus are the circles with equations $|z|=\text{const}$ (see Fig. 20). Function (27) is analytic in the two-connected region $z\neq 0$ (i. e. in the complex plane with deleted origin), and when the variable point makes a circuit about the point $z=0$ the function w gains the increment $\Delta w=\frac{\Gamma}{2\pi}\cdot 2\pi=\Gamma$. Consequently, at this point the circulation Γ is generated (the case shown in Fig. 20 corresponds to $\Gamma>0$). The point $z=0$ is said to be a **vortex point** for potential (27), and the corresponding field **A** is called a **vortex element.** We have

$$\frac{dw}{dz}=\frac{\Gamma}{2\pi i}\frac{1}{z}=\frac{\Gamma}{2\pi i}\frac{1}{x+iy}=\frac{\Gamma}{2\pi}\frac{-y-ix}{x^2+y^2}$$

and therefore, by Sec. 14, the field **A** is expressed as

$$\mathbf{A}=\frac{\Gamma}{2\pi(x^2+y^2)}(-y\mathbf{i}+x\mathbf{j}) \tag{28}$$

There is a curious thing here: the field of the vortex element is irrotational for $z\neq 0$. If **A** is interpreted as the velocity field of a plane-parallel flow of a fluid (in this case it is more natural to speak of an **axis of vortex** or **axis of whirl** described by the equations $x=y=0$), the infinitesimal volumes of the fluid

undergo translation and deformation (this question is discussed in more detail in Sec. V.1.5) but their angular speeds are equal to zero.

If we consider this field as being defined in the whole xy-plane including the point $z=0$ then reasoning as in Sec. I.2.4 we conclude that

$$\operatorname{div}\mathbf{A}=0,\ \operatorname{rot}\mathbf{A}=\Gamma\delta(x)\,\delta(y)\,\mathbf{k}$$

and hence the first condition (26) is violated at the origin. Therefore if we want to apply to this case what was said in Sec. 14 we must delete the point $x=y=0$.

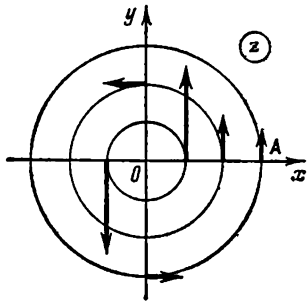


Fig. 20

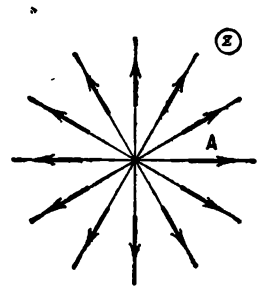


Fig. 21

A similar investigation of the complex potential

$$w=\frac{Q}{2\pi}\operatorname{Ln}z \tag{29}$$

where Q is real leads to the point source shown in Fig. 21 (for the case $Q>0$ this field was considered in Sec. I.2.4).

Combining potentials (27) and (29) we can obtain some other fields encountered in various applications. For example, the complex potential of a dipole (see Sec. I.2.4) with moment M located at the point $z=0$ and having the x-axis as its axis is equal to

$$w=\lim_{h\to 0}\left(\frac{Mh^{-1}}{2\pi}\operatorname{Ln}(z-h)+\frac{-Mh^{-1}}{2\pi}\operatorname{Ln}z\right)=-\frac{M}{2\pi}\cdot\frac{1}{z} \tag{30}$$

If a source of a field is distributed along a curve (L) with linear density $q=q(\zeta)$ $(\zeta\in(L))$ the corresponding field is referred to as a field of a **simple layer** and its potential is

$$w=\frac{1}{2\pi}\int_{(L)}q(\zeta)\operatorname{Ln}(z-\zeta)\,|d\zeta|$$

where $|d\zeta|$ is the element of arc length of the curve (L). If vortex (27) is similarly distributed along an arc (L) we speak of

a **vortex line**. A dipole of type (30) distributed along a line in such a way that its axis coincides with the normal to the line at each point is called a **double layer**. The corresponding potentials are expressed by the formulas

$$w=\frac{1}{2\pi i}\int\limits_{(L)}\gamma(\zeta)\,\mathrm{Ln}\,(z-\zeta)\,|\,d\zeta\,|$$

and

$$w=\frac{i}{2\pi}\int\limits_{(L)}\frac{m(\zeta)}{z-\zeta}\,d\zeta$$

whose derivation is left to the reader.

Let us consider another simple example. Take the potential

$$w=z^2=x^2-y^2+i2xy \tag{31}$$

whose stream lines are the hyperbolas $xy=\text{const}$ (Fig. 22). The corresponding family of vector lines has a singular point of the type of a saddle (e. g. see [107], Sec. XV.7). To find the direction of the field along these lines it is sufficient to construct the vector of the field at an arbitrary non-singular point and then continuously extend it (since the directions of the field at points which are close to one another are also close). For instance, we have

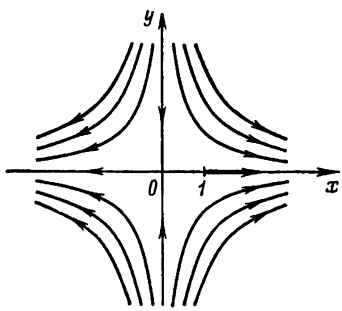


Fig. 22

$$\left(\frac{dw}{dz}\right)^*\Big|_{z=1}=(2z)^*\Big|_{z=1}=2$$

and therefore the directions of the field are those shown in Fig. 22 (check it up!).

Every field can be considered not in the entire z-plane but in some region of the plane. For instance, the field with potential (31) regarded as defined only in the quadrant $x\geqslant 0$, $y\geqslant 0$ can be interpreted as the velocity field of an ideal liquid flowing around the interior of a right angle.

16. Boundary-Value Problems and Conformal Mappings. As was shown in Sec. 14, the construction of a field **A** satisfying conditions (26) in a given plane region (G) is connected with a considerable degree of arbitrariness because we can take as the potential of the field any analytic function $w(z)$ (which can be multiple-valued in the general case) defined in (G) and having a single-valued derivative. Therefore, to specify a concrete field we must set some additional conditions guaranteeing the solvability of the problem and the uniqueness of the solution. These condi-

tions are usually taken in the form of scalar equations, called **boundary conditions**, connecting the projections of the sought-for vector field (the derivatives of the projections can also be involved) at each point of the boundary (S) of the domain (G). (Therefore in this case a boundary condition is a relation connecting functions but not numbers as in the theory of ordinary differential equations; e. g. see [107], Sec. XV.16). A **boundary-value problem**, i. e. a problem of determining a field for given boundary conditions, may also include several subsidiary relations connecting numbers.

For example, let the domain (G) be finite (in this case we speak about an **interior** boundary-value problem) and simply connected and the field $\mathbf{A}$ be interpreted as the velocity field of a flow of an ideal liquid. From the point of view of physics it then appears natural that this field will be uniquely determined if at each point $M \in (S)$ the intensity of flux through the contour (S) is given, i. e. if we set the values of the function

$$A_n|_M = g(M) \quad (M \in (S)) \tag{32}$$

where, for definiteness, the direction n is taken along the outer normal to (S). Besides, the condition

$$\int_{(S)} g(M)\, dS = 0 \tag{33}$$

must also hold (why?). The fact that the problem of constructing a field $\mathbf{A}$ satisfying equations (26) and boundary condition (32) (under the necessary condition (33)) has exactly one solution can be given a purely mathematical proof without references to physics.

If the domain (G) is finite but k-connected $(k \geqslant 2)$, there can appear circulatory motions around its lacunas (see Sec. 14). Therefore we must additionally set the values

$$\oint_{(L_j)} A_\tau\, dL = \Gamma_j \quad (j = 1,\ 2,\ \ldots,\ k-1)$$

for closed contours (L_1), (L_2), $\ldots$, (L_{k-1}) each of which envelopes only one of the lacunas of the domain (G). If the domain (G) is obtained from the plane by deleting one or several finite regions (in this case we speak about an **exterior** boundary-value problem) the physical considerations again imply that we must additionally set the value $\mathbf{A}_\infty$ of the field at infinity ("the velocity of the incident flow of the fluid"). This condition can also be given a purely mathematical justification. But it should be noted that in a more precise mathematical formulation of the problem it is required that the field should have no singularities of the type of a dipole (see Sec. 15) on the boundary (S) and also no singularities with a higher order of growth of the field.

There are many other types of boundary condition. Various boundary problems are treated in detail in the theory of equations of mathematical physics. Here we shall only discuss some simple facts which are directly related to the subject matter of the present chapter.

For definiteness, let us consider the exterior boundary-value problem of determining a plane field A satisfying conditions (26) outside a closed contour (S) and the boundary and subsidiary conditions

$$A_n|_{(S)} \equiv 0, \quad \mathbf{A}|_{z=\infty} = \mathbf{A}_\infty, \quad \oint_{(L)} A_\tau dL = \Gamma \tag{34}$$

where the values $\mathbf{A}_\infty$ and Γ are given. The solution of this problem can be interpreted as the velocity field of a flow of an ideal liquid incident on an impenetrable contour (S) with velocity $\mathbf{A}_\infty$, the circulation Γ of the field over a contour (L) (which encloses (S)) being known (see Fig. 23).

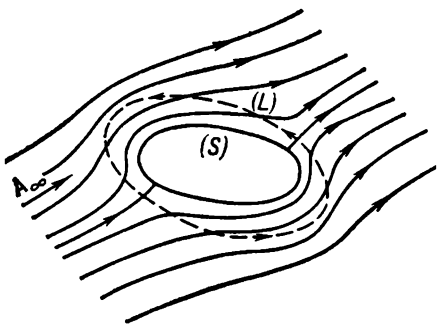


Fig. 23

Let us construct the complex potential $w(z)$ of the field A (see Sec. 14). Conditions (34) can be rewritten in the form

$$\left.\frac{\partial \varphi}{\partial n}\right|_{(S)} \equiv 0, \quad \left.\frac{dw}{dz}\right|_{z=\infty} = a, \quad \Delta_L w = \Gamma \tag{35}$$

where a is known (why?). The first condition (35) is a functional relation which can be written as $\left.\frac{\partial \psi}{\partial \tau}\right|_{(S)} \equiv 0$, i. e. $\operatorname{Im} w|_{(S)} \equiv \text{const}$. This means that the whole contour (S) is a portion of a stream line (which is also evident from the physical considerations).

The method of solving the above boundary-value problem with the aid of conformal mappings is based on the following idea. Suppose that we manage to construct a conformal mapping $\zeta = f(z)$ of the domain (G) (the exterior of (S)) onto a domain (H) of the ζ-plane, which is the exterior of a closed curve (T). Under the mapping, the points of the contour (S) correspond to those of the line (T), and the point $z = \infty$ corresponds to the point $\zeta = \infty$. Imagine that the line (T) is so simple that for the domain (H) we can easily solve a similar problem of constructing an analytic function $\omega(\zeta)$ satisfying the conditions

$$\operatorname{Im} \omega|_{(T)} = \text{const}, \quad \left.\frac{d\omega}{d\zeta}\right|_{\zeta=\infty} = b, \quad \Delta_\Lambda \omega = \Gamma \tag{36}$$

where the value of the constant b will be specified later and (Λ) is a contour lying in (H) and enveloping (T). Let us then "transfer" the values of ω from the points of (H) to the corresponding points of (G), i. e. put $\omega(z)=\omega(f(z))$. It can readily be shown (do it!) that this function is analytic in (G) and satisfies the first and the third conditions (35). Besides, we have

$$\left.\frac{dw}{dz}\right|_{z=\infty}=\left.\frac{d\omega}{d\zeta}\right|_{\zeta=\infty}\cdot\left.\frac{d\zeta}{dz}\right|_{z=\infty}=bf'(\infty)$$

and hence the second condition (35) is also satisfied if we put $b=\frac{a}{f'(\infty)}$.

Let us consider an example. Suppose that (S) is the circle $|z|=R$, a is a real number and $\Gamma=0$. By Sec. 11, the function $\zeta=\frac{c}{2}\left(\frac{z}{R}+\frac{R}{z}\right)$ maps conformally the domain (G) (which is the exterior of (S)) onto the domain (H), the exterior of the line segment (T) with end-points $\zeta=-c$ and $\zeta=c$. For this domain, boundary-value problem (36) with $b=\frac{a}{f'(\infty)}=\frac{2aR}{c}$ has a simple solution: $\omega=\frac{2aR}{c}\zeta$ (check it up!). Transferring these values to (G) we obtain in the z-plane the solution

$$w=\frac{2aR}{c}\frac{c}{2}\left(\frac{z}{R}+\frac{R}{z}\right)=a\left(z+\frac{R^2}{z}\right)$$

which is a potential of a non-circulatory flow around the circle. Now it is easy to find the sought-for potential of the circulatory flow around the circle:

$$w=a\left(z+\frac{R^2}{z}\right)+\frac{\Gamma}{2\pi i}\operatorname{Ln} z$$

(check it up!). Transferring the values of this potential to the points of other domains we can obtain the solution of problem (35) for various contours (S).

The application of conformal mappings to interior boundary-value problems is similar to the above.

Conformal mappings are also applicable to an interesting class of problems which has not yet been thoroughly investigated: these are the problems in which a portion of the boundary of the domain (G) or the whole boundary is not known beforehand. As an example, let us discuss the general scheme of solving one of such problems, namely the problem of a free outflow of a weightless ideal liquid. Consider the process of outflow depicted in Fig. 24 where the positions of the rigid walls (shown in heavy lines) are given but the shape of the free surface ae of the outflowing liquid is not known in advance. The points b, c, d and e are thought

of as being at infinity, and hence $z = \infty$ is a *double point* of the boundary of the domain (G) in the sense that at the points of the vessel lying at "infinity" there is a source of the liquid and at the points of the outflowing liquid lying "at infinity" there is a sink.

For the portion of the boundary whose shape is not set beforehand we must introduce a subsidiary boundary condition. If **A** is the velocity vector then we have the condition $A_n = 0$ on the

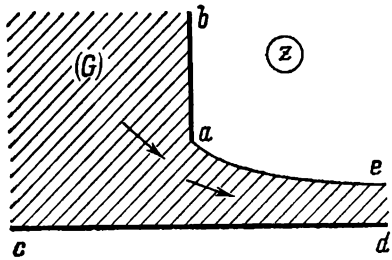


Fig. 24

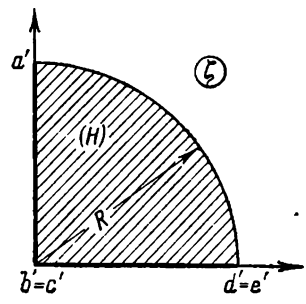


Fig. 25

whole boundary, and the subsidiary condition mentioned above is expressed by the relation $|\mathbf{A}| = \text{const}$ which is obviously implied by the law of conservation of energy. To construct the curve *ae* and the complex potential $w(z)$ of the field **A**, let us consider the mapping of the domain (G) into the ζ-plane determined by the formula $\zeta = w'(z)$. Taking into account the meaning of the derivative (Sec. 14) and the boundary conditions for the field **A**, we see that the image of the domain (G) under the mapping is a quarter-circle, which is a domain of type (H), whose boundary is known (see Fig. 25 where the points corresponding to the indicated points of the contour (G) are also shown). Virtually, under this mapping the line *cd* goes into the segment $c'd'$ of the real axis, with the origin included (why?), since we have $A_x > 0$ and $A_y = 0$ on *cd*, and the line *ba* goes into the segment $b'a'$ of the axis of imaginaries because we have the boundary conditions $A_x = 0$, $A_y < 0$ on *ba*.

The problem of constructing field (26) in the domain (H) which satisfies the boundary condition $A_n = 0$ and has a source of given strength Q at c' and a sink of the same strength at d' is readily solved by mapping the domain (H) on another domain of a simpler form, but we shall not consider this question here. It turns out that the corresponding complex potential $\omega(\zeta)$ is equal to $-\frac{2Q}{\pi} \operatorname{Ln}\left(\frac{\zeta}{R} - \frac{R}{\zeta}\right)$. Transferring its values to (G) we obtain the

potential $w=\omega(\zeta(z))$ of the original problem. But we have $\zeta=\frac{dw}{dz}$, and thus we arrive at the equality $w=\omega\left(\frac{dw}{dz}\right)$ which is a differential equation with the unknown function $w(z)$. Thus, to complete the solution of the problem we must solve the differential equation, and only after the equation has been solved we obtain the sought-for potential. (What has been said is a characteristic feature of a *reverse problem*; the problem in question can be, in a certain sense, regarded as a reverse of an ordinary boundary-value problem since a part of the boundary is not known). If the potential is found the required line *ae* is readily determined as a stream line passing through the given point *a*. The corresponding calculations (which we leave to the reader) yield, for the curve *ae*, the equation

$$z=h_0 i-\frac{2Q}{\pi}\ln\frac{(e^{it}-i)(1+i)}{(e^{it}+i)(1-i)}+\frac{2Q}{\pi}\frac{i}{R}(e^{it}-1)$$

where h_0 is the width of the vent, and the parameter t varies from 0 to $\frac{\pi}{2}$. Passing to the limit for $t\longrightarrow\frac{\pi}{2}$ we obtain the value $h_\infty=h_0-\frac{2Q}{\pi R}$ of the width of the stream at infinity. But we have $h_\infty R=Q$ (why?), whence $R=\frac{Q}{h_0}\left(1+\frac{2}{\pi}\right)$. The parameters h_0 and Q remain arbitrary.

17. General Remarks on Conformal Mappings. In this section we shall present, without proofs, some general facts concerning conformal mappings. These proofs are rather complicated; they can be found in the books mentioned at the beginning of the chapter.

Let us begin with the problem of possibility (we mean the possibility in principle, irrelative of the technical performance) of constructing a conformal mapping of an arbitrary domain (G) onto any given domain (H). *In what follows we suppose that the domains are open and do not contain the point at infinity.* G.F.B. Riemann proved in 1851 that *if the domains* (G) *and* (H) *are simply connected and do not coincide with the entire complex plane a mapping of this type always exists.* In constructing such a mapping we have three real degrees of freedom. In particular, if we set the image in (H) of one of the points of (G) (this specifies two real parameters) and the angle of rotation of an infinitesimal neighbourhood of this point (this determines the value of one more parameter), the mapping in question is uniquely specified. For the same purpose we can set the images of one interior point and one boundary point of the domain (G) or the images of three boundary points (in the latter case the images on the oriented

boundary of (H) must be placed in the same order as their preimages on the boundary of (G)).

The situation becomes more complicated in the case of multiply connected domains. First of all, domains with different connectivity numbers cannot be conformally mapped on each other (see Sec. I.2.2.). Moreover, they even cannot be topologically equivalent (Sec. 7). For instance, a simply connected domain (G) cannot be mapped conformally onto a domain (H) which is not simply connected: if a mapping of that type existed, (S) were a closed contour in the domain (H) enveloping one of its lacunas and (L) were the corresponding contour lying in (G), we could contract (L) to a point and would arrive at a contradiction by considering the corresponding deformation of (S) (why?).

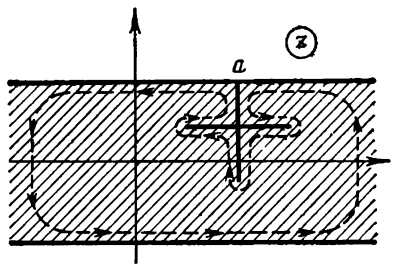


Fig. 26

Even the domains with the same connectivity number by far not always can be conformally mapped onto one another. For instance, it can be proved that every 2-connected domain other than the complex plane with a deleted point can be mapped conformally on the annulus $R_0 < |z| < 1$ where $0 \leqslant R_0 < 1$ and the number R_0 is uniquely determined. Thus, two ring-shaped domains with different values of R_0 cannot be conformally mapped onto each other. For the domains with higher connectivity numbers the situation is still more complicated and we shall not dwell on this problem here.

When an open domain (G) is conformally mapped on another open domain (H) the boundary points of the domains are in one-to-one correspondence as well. (This property can also be expressed as follows: every conformal mapping of (G) onto (H) can be continued to obtain a homeomorphism of the corresponding closed regions.) It should also be taken into account that a boundary point which can be approached along essentially different paths within the domain (or, more precisely, along paths which cannot be continuously transformed into one another in a sufficiently small neighbourhood of the point without leaving the domain) is regarded as a **multiple point** of the boundary, that is as several points, depending on the number of those paths. For example, consider the simply connected domain which is an infinite strip with a cross-shaped cut (see Fig. 26). All the boundary points of this domain except the point $z = \infty$ and the points of the cut are of multiplicity 1, i.e. **simple**. All the points of the cross-shaped cut except its centre are **double points** (of multiplicity 2) whereas

the centre of the cross is of multiplicity 4. The dotted line in Fig. 26 indicates the order in which the boundary points of this domain are mapped on the corresponding points of the circumference of a circle when the domain itself is mapped conformally on this circle. In particular, there are four different points on the circumference which correspond to the centre of the cross-shaped cut.

In the general case a conformal mapping remains conformal on the boundaries of the domains except separate points. Namely, let a be a boundary point of the domain (G) and a' the corresponding point on the boundary of (H). Let both points be finite. Then if a and a' are not corner points of the boundaries and the curvatures of the boundaries are finite at these points (the latter requirement may be considerably weakened), the mapping is conformal at a and a'. But if the boundary of the domain (G) has a corner point with an angle of magnitude α at the point a and the boundary of (H) forms an angle α' at the point a', the conformal mapping of (G) onto (H) is expressed, in an infinitesimal neighbourhood of the point a, by the formula $\zeta = a' + C(z-a)^{\frac{\alpha'}{\alpha}}$ accurate to within infinitesimals of higher order (cf. Sec. 9). We also note that in some cases there appears an additional factor of the type $[\mathrm{Ln}\,(z-a)]^{\beta}$ in the second summand. We see that if $\alpha \neq \alpha'$ the mapping is no longer conformal at the point a.

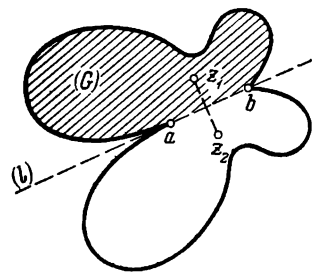


Fig. 27

In studying conformal mappings it is sometimes advisable to apply the so-called **symmetry principle**. Suppose that the boundary of the domain (G) contains a rectilinear segment ab and that the whole domain (G) lies on one side of the straight line (l) which is the extension of this segment. Let the domain (H) possess the same properties relative to a rectilinear part $a'b'$ of its boundary and its extension (l'). Suppose that there is a conformal mapping $\zeta = f(z)$ of the domain (G) onto the domain (H) such that the line segments ab and $a'b'$ correspond to each other. Now we "double" the domain (G) by adding to it its reflection through (l) and the portion ab of its boundary (see Fig. 27 where the original domain (G) is shaded). Denote the "doubled" domain as $(\tilde{G})$ and perform a similar duplication of the domain (H) to obtain the doubled domain $(\tilde{H})$. We can now easily extend (continue) the mapping f to obtain a conformal mapping $\tilde{f}$ of $(\tilde{G})$ onto $(\tilde{H})$: for this purpose we simply put $\tilde{f}(z) = f(z)$ if the point z is in (G) or on its boundary, and for every pair of points z_1 and z_2 symmetric

with respect to (l) (see Fig. 27) we assume that the points $\tilde{f}(z_1)$ and $\tilde{f}(z_2)$ are symmetric with respect to (l') (where $\tilde{f}(z_1)$ or $\tilde{f}(z_2)$ is of course equal to the corresponding value $f(z_1)$ or $f(z_2)$ depending on which of the points z_1 and z_2 is in (G)). Let the reader show that the mapping $\tilde{f}$ thus constructed is conformal in the reflection of (G) and hence is conformal throughout the domain $(\tilde{G})$.

A similar extension of a conformal mapping can be carried out, on the basis of symmetry, across an arc of a circle (see Sec. 8). Moreover, it turns out that an analogous continuation can be made across an arbitrary arc without singular points if we "straighten" this arc beforehand by means of an additional conformal mapping. But in the general case it is rather difficult to indicate wider domains to which a given mapping can be continued, and it is clearly impossible to extend a nonlinear mapping to the whole complex plane.

There are some results obtained by various authors in studying the dependence of a conformal mapping on variations of the boundary of its domain of definition. These results sometimes make it possible to obtain certain characteristics of a mapping of a domain with a complicated boundary by embedding it into a region between two simpler boundaries for which the mappings are easily found. Here we shall limit ourselves to formulating **Lindelöf's theorem** on variations of a conformal mapping of a domain on a circle when the sizes of the domain are reduced (see below). For the proof and more detail concerning these problems we refer the reader to [81].

Suppose that a simply connected domain (G) with boundary (S) is entirely contained within a finite simply connected domain $(H) \neq (G)$ with boundary (T). Let a function $\chi(z)$ determine a conformal mapping of (G) on unit circle and a function $\psi(z)$ specify a conformal mapping of the domain (H) on the same circle. Let also $0 \in (G)$ and $\chi(0) = \psi(0) = 0$. Then it turns out that for any $r\,(0 < r < 1)$ the closed contour $|\chi(z)| = r$ is strictly contained (without tangency!) inside the closed contour $|\psi(z)| = r$, and $|\chi'(0)| > |\psi'(0)|$. If (S) and (T) have at least one common point z_0 then $|\chi'(z_0)| < |\psi'(z_0)|$. If the curves (S) and (T) have, respectively, polar equations $\rho = \alpha(\varphi)$ and $\rho = \beta(\varphi)$ with single-valued right-hand sides, and $\beta(\varphi_0) - \alpha(\varphi_0) = \max\limits_{\varphi} [\beta(\varphi) - \alpha(\varphi)]$, then $\alpha(\varphi_0)\,|\chi'(\alpha(\varphi_0)\,e^{i\varphi_0})| \geqslant \beta(\varphi_0)\,|\psi'(\beta(\varphi_0)\,e^{i\varphi_0})|$.

It should also be noted that when a small portion of the boundary of a domain which is mapped conformally is changed the influence of this variation rapidly decreases when the boundary point is moved away from that portion.

18. Application of Small-Parameter Method. In practical calculations it is possible to obtain an exact conformal mapping of

a given domain onto another domain only when the domains are of a comparatively simple shape. Therefore, a number of approximate methods of conformal mapping were developed which are applicable to wider classes of domains. These methods can be found in [81] and, particularly, in [37], [58], [66] and [87]. Here we shall limit ourselves to the application of the small-parameter method (e.g. see [107], Sec. V.5) to constructing a conformal mapping of a domain which is close to a circle on that circle. This result can be applied to a more general problem of determining a conformal mapping of a domain obtained by a small variation of a domain (G) onto a domain (H) if the domains (G) and (H) are of a simple form. Indeed, let $g(z)$ and $h(z)$ be functions which determine, respectively, conformal mappings of the domains (G) and (H) on unit circle. Consider a domain (G_1) obtained from (G) by a small variation of its contour. Then the domain $g(G_1)$ is close to the unit circle (why?). If a function f_1 performs a conformal mapping of this domain onto the unit circle then the function $h^-(f_1(g(z))$ (where h^- is the inverse function of the function h) determines a conformal mapping of (G_1) on (H).

For our further aims we need the following

Lemma. *For a function $f(t)$ which can be expanded into a complex Fourier series* (e.g. see [107], Sec. XVII.26)

$$f(t)=\sum_{n=-\infty}^{\infty} c_n e^{int} \tag{37}$$

to have a constant modulus it is necessary and sufficient that

$$\sum_{n=-\infty}^{\infty} c_n^* c_{n+p}=0 \quad (p=1,\ 2,\ 3,\ \ldots) \tag{38}$$

If this condition holds then

$$|f(t)|^2 \equiv \sum_{n=-\infty}^{\infty} |c_n|^2 \tag{39}$$

To prove the lemma we pass in (37) to conjugate complex quantities, which results in

$$f^*(t)=\sum_n c_n^* e^{-int}=\sum_m c_m^* e^{-imt}$$

and multiply (37) by the latter relation to obtain

$$|f(t)|^2=\sum_{n,\,m} c_n c_m^* e^{i(n-m)t}=|\,n=m+p\,|=\sum_{m,\,p} c_{m+p} c_m^* e^{ipt}=$$
$$=\sum_m c_m c_m^*+\sum_{p\neq 0}\left(\sum_m c_{m+p} c_m^*\right) e^{ipt}$$

whence if follows that condition (38) holds for $p\neq 0$, and equality (39) is fulfilled. But we have $\sum_n c_n^* c_{n-p}=\left(\sum_n c_n^* c_{n+p}\right)^*$ (check it

up!), and therefore it is sufficient if condition (38) only holds for $p > 0$. The lemma has been proved.

Now we proceed to apply the small-parameter method. Let a domain (G_α) be bounded by a closed contour whose equation is

$$z = e^{it} + \alpha \sum_{n=-\infty}^{\infty} d_n e^{int} \tag{40}$$

where α is a real small parameter, and the sum it is multiplied by represents a periodic function with period 2π (which is necessary for the contour to be closed). For $\alpha = 0$ the domain (G_0) is a unit circle, and therefore we can start from the identity mapping $\zeta = z$. Following the scheme of the small-parameter method, let us suppose that for $\alpha \neq 0$ the sought-for mapping of the domain (G_α) on a circle (H) is of the form

$$\zeta = z + \alpha g_1(z) + \alpha^2 g_2(z) + \dots \tag{41}$$

To normalize the mapping let us assume that for $\alpha = 0$ we have

$$\zeta\Big|_{z=0} = 0, \quad \frac{d\zeta}{dz}\Big|_{z=0} = 1 \tag{42}$$

Here we have two relations specifying four real parameters (why?) while a conformal mapping, as was shown in Sec. 17, is completely characterized by setting the values of three independent real parameters. Therefore, in order to introduce an additional degree of freedom, we shall not fix the radius R of the circle (H) and simply put $|\zeta| < R$.

Let us try to construct the functions $g_j(z)$ as a power series. Formulas (41) and (42) indicate that there must not be constant and linear terms in these series, that is we have

$$\zeta = z + \alpha \sum_{k=2}^{\infty} a_{1k} z^k + \alpha^2 \sum_{k=2}^{\infty} a_{2k} z^k + \dots \tag{43}$$

The problem is now reduced to determining the coefficients a_{jk}.

The boundary of the domain (G) is mapped onto the circumference $|\zeta| = R$, and therefore the function of t which is obtained by substituting expression (40) into the right-hand side of (43), that is the expression

$$e^{it} + \alpha \sum_{n=-\infty}^{\infty} d_n e^{int} + \alpha \sum_{k=2}^{\infty} a_{1k} \left[e^{it} + \alpha \sum_{n=-\infty}^{\infty} d_n e^{int} \right]^k + \\ + \alpha^2 \sum_{k=2}^{\infty} a_{2k} \left[e^{it} + \alpha \sum_{n=-\infty}^{\infty} d_n e^{int} \right]^k + \dots \tag{44}$$

must have the constant modulus R. Hence, taking into account that all the coefficients c_n (which are obtained by regrouping the

terms entering into expression (44)) are series in powers of α, we can apply the above lemma. Therefore, each of the equalities (38) generates a sequence of relations obtained by equating to zero the coefficients in the powers of α. Let us write down initial terms of the expansions of the coefficients c_n:

$$c_n = \alpha d_n + \ldots \ (n \leqslant 0); \quad c_1 = 1 + \alpha d_1 + \ldots;$$
$$c_n = \alpha (d_n + a_{1n}) + \ldots (n \geqslant 2) \tag{45}$$

Substituting these expressions into (38) and equating to zero the coefficients in α we obtain (check it up!) the relation $(d_{p+1} + a_{1,p+1}) + d^*_{1-p} = 0$, whence

$$a_{1k} = -d_k - d^*_{2-k} \quad (k = 2, 3, 4, \ldots) \tag{46}$$

To obtain the coefficients a_{2k} we must write expansions (45) up to the terms with α^2 inclusive and take the coefficient in α^2 in expansion (38); we leave these calculations to the reader. It should also be noted that (38) and (45) yield initial terms of the expansion of R^2:

$$R^2 = 1 + 2 \operatorname{Re} d_1 \alpha + \ldots \tag{47}$$

If we wanted to construct a mapping of the domain (G_α) on the unit circle it would be sufficient to divide function (43) by the value of R thus found.

The linear part of the obtained expression can also be represented in the form of an integral which sometimes turns out to be more convenient. To derive this integral representation let us construct the corresponding influence function. Since the problem of determining a conformal mapping is non-linear, we must find the normalized result of a small action (on the notion of an influence function and on the last remark see, for example, [107], Sec. XIV.26). For this purpose we first suppose that the polar equation of the boundary of the domain (G_p) has the form $\rho = 1 + P\delta(\varphi - \varphi_0)$ $(0 \leqslant \varphi \leqslant 2\pi)$ where $|P| \ll 1$ and δ is the delta-function. This equation can also be written as $z = e^{i\varphi} + P\delta(\varphi - \varphi_0) e^{i\varphi}$, and so, comparing it with (40), we see that $t = \varphi$, $\alpha = P$ and $d_n = \frac{1}{2\pi} e^{(1-n)i\varphi_0}$ (check it up!). Let us consider (H) to be a unit circle and therefore divide function (43) by the value of R found from (47). Taking into account formula (46) we finally arrive at the influence function for the perturbation of the function specifying the conformal mapping:

$$G(z, \varphi_0) = \lim_{P \to 0} \frac{1}{P} \left(\frac{\zeta}{R} - z \right) -$$

$$- \lim_{P \to 0} \frac{1}{P} \left\{ \left[z - P \sum_{k=2}^{\infty} (d_k + d^*_{2-k}) z^k + \ldots \right] (1 - P \operatorname{Re} d_1 + \ldots) - z \right\} =$$

$$= -z\,\mathrm{Re}\,d_1 - \sum_{k=2}^{\infty}(d_k + d_{2-k}^{*})\,z^k =$$

$$= -\frac{1}{2\pi}\left(z + 2\sum_{k=2}^{\infty} e^{(1-k)i\varphi_0} z^k\right) = -\frac{1}{2\pi}\,\frac{1+e^{-i\varphi_0 z}}{1-e^{-i\varphi_0 z}}\,z$$

(let the reader check up the calculations). Now, following the general technique of applying an influence function, we find that if the contour of the domain to be mapped has a polar equation of the form $\rho = 1 + p(\varphi)$ where $|p(\varphi)| \ll 1$, then, to within infinitesimals of higher order, the sought-for mapping is expressed by the formula

$$w = z + \int_0^{2\pi} G(z;\ \varphi_0)\,p(\varphi_0)\,d\varphi_0 = z - \frac{1}{2\pi}\int_0^{2\pi} \frac{1+e^{-i\varphi z}}{1-z^{-i\varphi z}}\,z p(\varphi)\,d\varphi$$

On further results concerning these problems (in particular, conformal mappings of domains of a more complicated type) see [81].

§ 2. INTEGRATION AND POWER SERIES

1. Integral. Let (L) be an oriented curve in the complex z-plane. Suppose that at the points of the curve the values of a function $f(z)$ are given. Then the definition of the integral

$$\int_{(L)} f(z)\,dz = \lim \sum_{k=1}^{n} f(\sigma_k)\,\Delta z_k \qquad (1)$$

is completely analogous to that of the real line integral with respect to a coordinate (e.g. see [107], Sec. XIV.23), the meaning of the notation being indicated in Fig. 28. If the line (L) is of a finite length and the function $f(z)$ assumes at the points $z \in (L)$ finite values, integral (1) takes on a finite complex value as well. But if at least one of these conditions is violated the integral $\int_{(L)} f(z)\,dz$ is regarded as *improper*, and the question of its convergence is treated as in the case of real integrals.

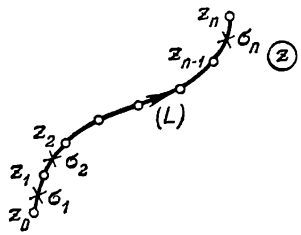


Fig. 28

Integral (1) is readily reduced to real line integrals. Indeed, let us put $f(z) = u(z) + iv(z)$, $z = x + iy$. Then we have

$$\int_{(L)} f\,dz = \int_{(L)} (u+iv)(dx+i\,dy) = \int_{(L)} (u\,dx - v\,dy) + i\int_{(L)} (v\,dx + u\,dy) \qquad (2)$$

Therefore, integral (1) possesses the same simple properties as line integrals with respect to a coordinate. In particular, it should be mentioned that the change of the orientation of the path (L) results in the multiplication of integral (1) by -1.

If it is possible to find an antiderivative $F(z)$ of the function $f(z)$, integral (1) can be computed by the *Newton-Leibniz formula* which in this case is written in the form

$$\int_{(L)} f(z)\,dz = F(z)\big|_{(L)}$$

where the symbol on the right-hand side designates the increment of the function $F(z)$ when the point z describes the line (L) in accordance with its orientation. If the function $F(z)$ is multiple-valued (the function $f(z)$ is always supposed to be one-valued) then, when applying the formula, we must begin with one of its values and follow its continuous variation.

Taking into account that the modulus of a sum does not exceed the sum of the moduli of the summands, and $|dz| = dL$ (the latter is the element of arc length) we obtain the following estimations for the integral:

$$\left|\int_{(L)} f(z)\,dz\right| \leqslant \int_{(L)} |f(z)|\,dL \leqslant \max_{z\in(L)} |f(z)| \cdot L \tag{3}$$

where, as usual, L denotes the length of the arc (L).

2. Integral of an Analytic Function. *Let the function $f(z)$ be single-valued and analytic at all points of an open domain (G).* Then we see that for the integrals on the right-hand side of (2) the condition of path-independence is fulfilled. In fact, for an integral of the form $\int (P\,dx + Q\,dy)$, this condition is written as $\frac{\partial P}{\partial y} \equiv \frac{\partial Q}{\partial x}$ (e.g. see [107], Sec. XIV.24). It also means that the field $\mathbf{A} = P\mathbf{i} + Q\mathbf{j}$ is irrotational. Applying this condition to integrals on the right-hand side of (2) we arrive at the Cauchy-Riemann equations (1.6) (check it up!) which hold for every analytic function.

Applying the results of Secs. I.2.1 and I.2.2 we immediately obtain the following assertions.

If the domain (G) is simply connected, then, for any fixed z_0, the integral

$$\int_{z_0}^{z} f(z)\,dz \tag{4}$$

is dependent solely on z and is independent of the path of integration lying in (G). This integral is an antiderivative of $f(z)$. *For*

any closed contour (L) *in* (G) *the equality*

$$\oint_{(L)} f(z)\,dz = 0 \tag{5}$$

holds.

If the domain (G) *is multiply connected, then, in the general case, integral* (4) *is an infinite-valued function of* z. The branches of this function differ from one another by constant summands and serve as antiderivatives of $f(z)$. *Equality* (5) *is guaranteed for any closed contour bounding a region entirely contained in* (G). For other closed contours (L) we can only assert that *the integral* $\oint_{(L)} f(z)\,dz$ *remains invariable if the path* (L) *is deformed continuously without leaving the domain* (G). (The latter assertion is also true for non-closed contours if, in the process of deformation, their end-points are kept fixed.)

In particular, we conclude that *if* $f(z)$ *is a single-valued analytic function everywhere inside an arbitrary closed contour* (L) *and on it, the integral of* $f(z)$ *over* (L) *satisfies relation* (5). This important assertion is known as **Cauchy's integral theorem.**

It should be noted that when we say that a function is "analytic on a contour" we usually mean that it is analytic in a curvilinear strip containing the contour. But in Cauchy's integral theorem and some other similar propositions it is sufficient to require that the function in question be analytic strictly inside the contour and continuous in the closed region bounded by it (including the points of the boundary).

We again stipulate that here and henceforward we only consider integrals of single-valued functions, and therefore in what follows we shall not mention this condition again. If, according to the original definition, a given function is multiple-valued, we shall integrate the concrete one-valued branch of the function we are interested in and suppose that the path of integration does not intersect the corresponding cuts. There are some more complicated cases when it is necessary to reject this condition but then we must stipulate what values of the integrand are meant and carefully consider the applicability of integral theorems.

3. Laurent Series. Laurent's series (named after P. M. H. Laurent, 1813-1854, a French mathematician) are widely used in the theory of analytic functions. A Laurent series is a generalization of a power series and is written in the form

$$\sum_{k=-\infty}^{\infty} c_k z^k = \ldots + \frac{c_{-2}}{z^2} + \frac{c_{-1}}{z} + c_0 + c_1 z + c_2 z^2 + \ldots \tag{6}$$

Thus, in contrast to an ordinary power series, here we also deal

with negative integral exponents. Therefore, generally speaking, Laurent's series involves the summation with respect to an index whose range extends from $-\infty$ to $+\infty$ (such series are referred to as **two-way series**). Accordingly, the series

$$c_0+c_1z+c_2z^2+\dots \quad \text{and} \quad \frac{c_{-1}}{z}+\frac{c_{-2}}{z^2}+\frac{c_{-3}}{z^3}+\dots \tag{7}$$

are termed the **regular** and **principal parts** of series (6).

The first series (7) is an ordinary power series and therefore it converges in a circle $|z|<R_2$ (e.g. see [107], Sec. XVII.14). The other series (7) is a power series in the variable $\frac{1}{z}$ and hence it converges in a domain $\left|\frac{1}{z}\right|<\rho$, i.e. $|z|>\frac{1}{\rho}=R_1$. For both series (7) to be convergent simultaneously the variable z must satisfy the relation $R_1<|z|<R_2$. Of course, only the case $R_1<R_2$ is of interest, and then the latter relation determines a ring-shaped domain (**the annulus of convergence**) (K) in the z-plane (see Fig. 29) at whose points series (6) converges. This series can also be convergent at some points of the circles $|z|=R_1$ and $|z|=R_2$, but for $|z|<R_1$ and $|z|>R_2$ it is divergent. In Fig. 29 we see the case when $R_1>0$ and $R_2<\infty$.

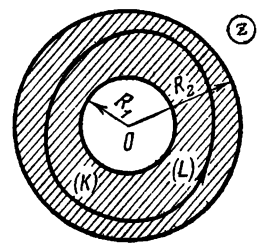


Fig. 29

According to the foregoing paragraph, the properties of ordinary power series (e.g. see [107], Sec. XVII.11) are in a natural manner extended to series (6). In particular, if the sum of series (6) in the annulus (K) is denoted by $f(z)$ then the formula

$$f(z)=\sum_{k=-\infty}^{\infty} c_kz^k, \quad z\in(K) \tag{8}$$

can be differentiated termwise any number of times and therefore *the sum of a Laurent series is an analytic function in the annulus where it converges.* If series (6) is integrated term-by-term we obtain the expression

$$F(z)=\left(\dots-\frac{c_{-3}}{2z^2}-\frac{c_{-2}}{z}+\text{const}+c_0z+\frac{c_1}{2}z^2+\frac{c_2}{3}z^3+\dots\right)+c_{-1}\operatorname{Ln}z \tag{9}$$

that is a sum of a Laurent series to which, in case $c_{-1}\neq 0$, the multiple-valued logarithmic function is added. Therefore, if we integrate both sides of formula (8) along a closed contour (L) lying in (K), enclosing the inner circle and making one circuit around it in the positive direction, we obtain, by applying the Newton-

Leibniz formula, the relation

$$\oint_{(L)} f(z)\,dz = F(z)\Big|_{(L)} = 0 + c_{-1}\cdot 2\pi i \tag{10}$$

Indeed, all the terms on the right-hand side of formula (9), except the last one, are single-valued analytic functions and therefore they receive zero increments when the point z traverses a closed contour.

A series of the form

$$\sum_{k=-\infty}^{\infty} c_k (z-z_0)^k$$

is also called Laurent's series; it converges in an annulus $R_1 < |z-z_0| < R_2$ with centre at the point z_0 *). Such series were studied by L. Euler as early as 1748.

4. Expanding an Analytic Function into a Laurent Series. In Sec. 3 we showed that the sum of Laurent's series is an analytic function in the annulus of its convergence. Here we shall prove that, conversely, *every function $f(z)$ which is one-valued and analytic in an annulus (K) can be expanded, in this annulus, into Laurent's series*

$$f(z) = \sum_{k=-\infty}^{\infty} c_k (z-z_0)^k,\ z \in (K) \tag{11}$$

where z_0 is the centre of the annulus. This important theorem was proved by P.M.H. Laurent in 1843.

For simplicity, let us suppose that the centre of the annulus is at the point $z=0$, i.e. the annulus is determined by the condition $R_1 < |z| < R_2$ where $0 \leqslant R_1 < R_2 \leqslant \infty$. Put $z = \rho e^{i\varphi}$ where $\rho = \text{const}$ and satisfies the inequalities $R_1 < \rho < R_2$. Then $f(\rho e^{i\varphi})$ is a periodic function of φ with period 2π (why?). The function $f(\rho e^{i\varphi})$ can be expanded in a Fourier series which we write in the complex form (e.g. see [107], Sec. XVII.26)

$$f(\rho e^{i\varphi}) = \sum_{k=-\infty}^{\infty} C_k(\rho)\, e^{ik\varphi} \tag{12}$$

where the coefficients (which are dependent on the fixed value of ρ) are determined by the formula

$$C_k(\rho) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\rho e^{i\varphi})\, e^{-ik\varphi}\, d\varphi =$$

$$= \frac{1}{2\pi} \int_{0}^{2\pi} f(\rho e^{i\varphi})\, e^{-ik\varphi}\, d\varphi \quad (k = 0,\ \pm 1,\ \pm 2,\ \ldots) \tag{13}$$

*) If $R_1 = 0$ and $R_2 < \infty$ the annulus reduces to a disk without its centre which is called a **deleted neighbourhood of z_0**, and then we have a **Laurent expansion at z_0**. If $R_1 > 0$ and $R_2 = \infty$ the annulus reduces to the exterior of a circle with the point at infinity deleted. Such a region is called a **deleted neighbourhood of ∞**, and in this case we have a **Laurent expansion at $z = \infty$**. — *Tr.*

To specify the dependence on ρ let us differentiate both sides of (13) with respect to ρ taking into account that ρ enters into the integrand as a parameter (e.g. see [107], Sec. XIV.20):

$$\frac{dC_k}{d\rho}=\frac{1}{2\pi}\int_0^{2\pi}\frac{\partial}{\partial\rho}\left[f(\rho e^{i\varphi})\,e^{-ik\varphi}\right]d\varphi=\frac{1}{2\pi}\int_0^{2\pi}e^{-ik\varphi}f'(\rho e^{i\varphi})\,e^{i\varphi}\,d\varphi$$

Now we put $e^{-ik\varphi}=u$, $f'(\rho e^{i\varphi})\,e^{i\varphi}\,d\varphi=dv$ and perform integration by parts using the expressions $du=-e^{-ik\varphi}ik\,d\varphi$ and $v=\frac{1}{i\rho}f(\rho e^{i\varphi})$ (check it up!):

$$\frac{dC_k}{d\rho}=\frac{1}{2\pi}e^{-ik\varphi}\frac{1}{i\rho}f(\rho e^{i\varphi})\Bigg|_0^{2\pi}+\frac{1}{2\pi}\int_0^{\pi}\frac{1}{i\rho}f(\rho e^{i\varphi})\,e^{-ik\varphi}ik\,d\varphi$$

The term $\frac{1}{2\pi}e^{-ik\varphi}\frac{1}{i\rho}f(\rho e^{i\varphi})\Big|_0^{2\pi}$ is equal to zero (why?), and therefore, comparing the above expression of $\frac{dC_k}{d\rho}$ with (13), we arrive at the differential equation

$$\frac{dC_k}{d\rho}=\frac{k}{\rho}C_k$$

whence

$$\frac{dC_k}{C_k}=k\frac{d\rho}{\rho},\quad C_k(\rho)=c_k\rho^k,\quad c_k=\text{const}$$

(check up the result!). Substituting the expression thus found into (12) and (13) we obtain the expansion

$$f(z)=\sum_{k=-\infty}^{\infty}c_k\rho^k e^{ik\varphi}=\sum_{k=-\infty}^{\infty}c_k(\rho e^{i\varphi})^k=\sum_{k=-\infty}^{\infty}c_k z^k$$

which holds for all $z\in(K)$. Hence, the Laurent theorem has been proved.

To derive an explicit formula for the coefficients c_k let us replace the summation index k on the right-hand side of (11) by l and divide both sides by $(z-z_0)^{k+1}$. This results in

$$f(z)\,(z-z_0)^{-k-1}=\sum_{l=-\infty}^{\infty}c_l\,(z-z_0)^{l-k-1}\qquad(14)$$

Now we take an arbitrary closed contour (L) lying in (K) and making one circuit about the point z_0 in the positive direction and integrate both members of (14) over this contour. When performing the integration we make use of the fact that for any integer $m\neq-1$ we have $\oint_{(L)}(z-z_0)^m\,dz=0$ while $\oint_{(L)}(z-z_0)^{-1}dz=2\pi i$ (why?). Therefore, after the integration of the right-hand side of (14) has been

carried out, only the term with $l=k$ remains nonzero and thus we obtain $\oint\limits_{(L)} f(z)(z-z_0)^{-k-1}\,dz = c_k \cdot 2\pi i$ whence

$$c_k = \frac{1}{2\pi i} \oint\limits_{(L)} \frac{f(z)}{(z-z_0)^{k+1}}\,dz \quad (k=0,\ \pm 1,\ \pm 2,\ \ldots) \tag{15}$$

Formula (11) can be valid outside the original annulus (K). For instance, consider the annulus (K) which is shaded in Fig. 30. Suppose that the function $f(z)$ is analytic throughout the complex z-plane except the points and lines of discontinuity marked by asterisks in Fig. 30. Then the above argument applies to the annulus with the same centre z_0 bounded by the circles (S_1) and (S_2) (Fig. 30), and therefore formula (14) is valid for this annulus. In Sec. 5 we shall prove that further extension of the annulus in which series (14) converges is impossible, that is this series is divergent inside (S_1) and outside (S_2). Thus, the circles bounding the maximal domain of convergence of a Laurent expansion of an analytic function pass through the points or lines of discontinuity of the function. We note that only a single-valued function can be expanded into Laurent's series. Of course we can also expand into a Laurent series a single-valued branch of a multiple-valued function. Then the circles bounding the domain of convergence of the expansion may also pass through the corresponding branch points lying on the lines of discontinuity of the branch.

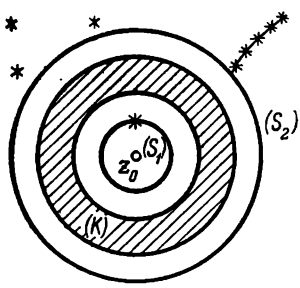


Fig. 30

Laurent's series of a given function is essentially dependent on the choice of the centre of the annulus in which the expansion is performed. For example, consider the function $f(z)=\frac{1}{z+1}+\frac{1}{z-2}$ which is analytic in the whole z-plane except the points $z=-1$ and $z=2$. Using the formula $(1-\alpha)^{-1}=1+\alpha+\alpha^2+\ldots(|\alpha|<1)$ we readily obtain the expansions

$$f(z)=\left(1-\frac{1}{2}\right)+\left(-1-\frac{1}{2^2}\right)z+\left(1-\frac{1}{2^3}\right)z^2+\left(-1-\frac{1}{2^4}\right)z^3+\ldots$$

$$(|z|<1)$$

$$f(z)=\ldots-\frac{1}{z^4}+\frac{1}{z^3}-\frac{1}{z^2}+\frac{1}{z}-\frac{1}{2}-\frac{1}{2^2}z-\frac{1}{2^3}z^2-\ldots(1<|z|<2)$$

and

$$f(z)=(1+1)\frac{1}{z}+(-1+2)\frac{1}{z^2}+(1+2^2)\frac{1}{z^3}+$$
$$+(-1+2^3)\frac{1}{z^4}+\ldots \quad (|z|>2)$$

(check up these expressions!).

5. Taylor Series. Let us consider a function $f(z)$ which is one-valued and analytic in a circle $|z-z_0|<R$ except possibly at its centre $z=z_0$. If the function is bounded in the vicinity of z_0, we have an expansion of type (11), with coefficients (15), in the "annulus" $0<|z-z_0|<R$. Taking the circle $|z-z_0|=\rho$, $0<\rho<R$, as the contour (L) and passing to the limit for $\rho\to 0$ we obtain, on the basis of estimate (3), the inequality

$$|c_k|\leqslant\frac{1}{2\pi}\max_{(L)}\left|\frac{f(z)}{(z-z_0)^{k+1}}\right|\cdot 2\pi\rho=\max_{(L)}|f(z)|\rho^{-k}$$

We see that for $k<0$ the right-hand side tends to zero as $\rho\to 0$ while the left-hand side is independent of ρ, and hence c_k is equal to zero. Thus, under the above assumptions, the Laurent series has no principal part and therefore it turns into an ordinary **Taylor series**. (We remind the reader that every power series is Taylor's series of its sum; e.g. see [107], Sec. XVII.13.)

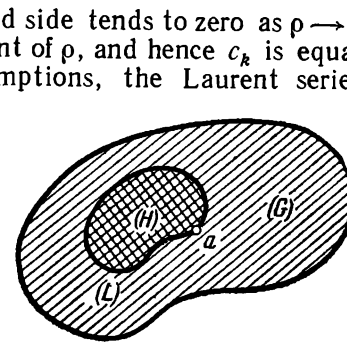


Fig. 31

In particular, it follows that *every single-valued analytic function can be expanded into Taylor's series in a neighbourhood of its regular* (i.e. non-singular) *point*. This result implies a number of corollaries. As is known (e.g. see [107], Sec. XVII.14), the sum of a power series possesses continuous derivatives of all orders. Thus we arrive at.

Corollary 1. *A function which is analytic in a domain possesses derivatives of all orders in this domain.* This is a curious result because the definition of analyticity given in Sec. 1.1 only requires the existence of the first derivative.

Corollary 2. *Let a function $f(z)$ be analytic in a domain (G), and let $f(z)\equiv 0$ in a subdomain (H) belonging to (G). Then $f(z)\equiv 0$ everywhere in (G).* Indeed, suppose the contrary, that is let there exist a closed curve (L) bounding a region in which $f(z)\not\equiv 0$ (Fig. 31), the function $f(z)$ being non-zero outside (L). Then we have $f^n(z)\equiv 0$ for all the derivatives in this region. The derivatives being continuous, we also have $f^n(z)=0$ for all n and all points z on the contour (L). Now, taking an arbitrary point $a\in(L)$

and expanding $f(z)$ in powers of $z-a$ we conclude that $f(z) \equiv 0$ everywhere in a neighbourhood of the point a which contradicts the hypothesis, and thus the corollary has been proved.

Corollary 3. Let a function $f(z)$ be analytic and not identically equal to zero. If $f(c)=0$ we say that the point c is a **zero (root)** of the function $f(z)$. The coefficients of the expansion of the function in powers of $z-c$ cannot all be equal to zero (why?). Therefore the expansion begins with a term $c_n(z-c)^n$ such that $c_n \neq 0$ and $c_k=0$ for $k<n$. The exponent n is called the **multiplicity (order)** of the zero $z=c$ of the function $f(z)$, i.e. of the equation $f(z)=0$ (e.g. see [107], Sec. VIII.8). The point $z=c$ is then referred to as a **root (zero) of multiplicity n, n-fold root** or **zero-point of order n.** In particular, if $n=1$, the zero is said to be **simple.** If otherwise, it is **multiple.** Thus we see that *all zeros of a single-valued analytic function* $f(z) \not\equiv 0$ *are of finite integral order* *). Zeros of non-integral or infinite order may only appear at branch points of a multiple-valued function (for instance, such as $f(z)=z^{\frac{3}{2}}$, $\varphi(z)=z^{\frac{1}{z}}$, etc.).

It should be noted that, generally speaking, for real functions of a real argument possessing derivatives (even for infinitely differentiable functions) neither of the above assertions holds (why?).

Now we can prove the assertion concerning the maximal annulus of convergence of a Laurent series which was formulated in Sec. 4. For this purpose we apply Corollary 2 to the difference between the right-hand and left-hand sides of formula (11). This difference is equal to zero in the shaded domain in Fig. 30 and therefore it is also equal to zero everywhere in the domain of convergence of series (11), and the function $f(z)$ is analytic (why?). If the series were convergent in a wider annulus than that shown in Fig. 30, the function $f(z)$ would not have singular points indicated in Fig. 30 because the sum of Laurent's series has no singular points inside the annulus of its convergence, which contradicts the hypothesis. The assertion has thus been proved.

In particular, it follows that the **radius of convergence** *of Taylor's series* (i.e. the radius of the **circle of convergence**) *corresponding to the expansion of an analytic function in powers of* $z-c$ *(where c is any regular point of the function) is equal to the distance from c to the nearest singular point of the function* $f(z)$. (It should be noted that if we deal with an expansion of a single-valued branch of a multiple-valued function its branch points are also regarded as its singular points.) In particular, if a function

*) A root of the equation $f(z)=A$, i.e. $f(z)-A=0$, is called an *A-point of* $f(z)$. Accordingly, the notions of a *simple A-point, multiple A-point, order of an A-point*, etc., are introduced by analogy with the corresponding notions related to a zero-point (which is an A-point for $A=0$).—*Tr.*

$f(z)$ does not have finite singular points the radius of convergence is infinite. In this case $f(z)$ is called an **entire** function. The above conclusion concerning the radius of convergence of Taylor's series of an analytic function makes it possible to determine this radius without investigating the remainder of the series.

Consider an example. Let us find the radius of convergence of the expansion of the function

$$\tanh z = \frac{\sinh z}{\cosh z} = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{16}$$

in powers of z. The numerator and denominator being entire functions, their quotient is a single-valued analytic function for all points z except those at which the denominator vanishes. Hence, function (16) has the singular points

$$z = \frac{\pi}{2} i + k\pi i \qquad (k = 0, \pm 1, \pm 2, \ldots)$$

(check it up!). The nearest singular points to the point $z = 0$ are $\pm \frac{\pi}{2} i$, and consequently the radius of convergence of the expansion is equal to $\frac{\pi}{2}$.

It should be noted that if we consider the argument of this function to be a real variable, that is if we take the function $\tanh x$ and investigate its expansion in powers of the real variable x, the coefficients of the Taylor series, and hence the radius of convergence as well, will be the same as in the case of the complex variable z. Therefore, the *interval of convergence* will be $-\frac{\pi}{2} < x < \frac{\pi}{2}$. The points $x = \pm \frac{\pi}{2}$ are not singularities of the function $\tanh x$, and therefore, from the point of view of the theory of real functions of a real variable, the fact that the series is divergent on every wider interval is inexplicable. The situation is only cleared if we pass to the complex variable z. Then it becomes evident that the imaginary singular points $z = \pm \frac{\pi}{2} i$ determine the *circle of convergence* $|z| < \frac{\pi}{2}$, and the above interval of convergence is the intersection of this circle with the x-axis (see Fig. 32).

Corollary 1 implies the so-called **Morera theorem** which is the converse of the Cauchy integral theorem (see Sec. 2): *if a function $f(z)$ is continuous in an open region (G), and formula* (5) *holds for every closed contour (L) lying in (G), the function is analytic in (G).* Indeed, the conditions of the theorem imply that integral (4) is path-independent, and if z_0 is a fixed point, the integral depends

solely on z, that is $\int_{z_0}^{z} f(z)\,dz = \Phi(z)$. But then we have $\Phi'(z) = f(z)$, i.e. $\Phi(z)$ is an analytic function. It possesses derivatives of all orders, and therefore the function $f(z) = \Phi'(z)$ also has a derivative and thus is analytic.

In particular, it follows from Morera's theorem that if a domain (G) is divided by a curve (S) into parts (G_1) and (G_2) (see

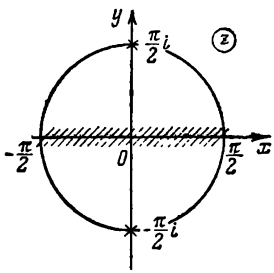


Fig. 32

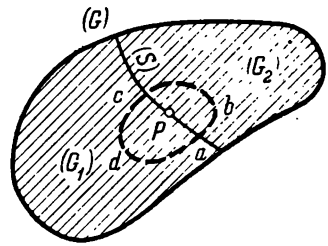


Fig. 33

Fig. 33), and there are analytic functions $f_1(z)$ and $f_2(z)$ defined, respectively, in (G_1) and (G_2), which assume on (S) the same values forming a continuous function on the line (S), the function

$$f(z) = \begin{cases} f_1(z) & \text{for } z \in (G_1) \\ f_2(z) & \text{for } z \in (G_2) \\ f_1(z) = f_2(z) & \text{for } z \in (S) \end{cases}$$

is analytic in (G). Virtually, this function is obviously continuous, and formula (4) for the contours of the type of *abcd* (Fig. 33) is implied by the equalities

$$\int_{abcda} f(z)\,dz = \int_{abcpa} f(z)\,dz + \int_{apcda} f(z)\,dz = \int_{abcpa} f_2(z)\,dz + \int_{apcda} f_1(z)\,dz$$

Consequently, an analytic function can have lines of discontinuity but there cannot exist curves on which the function itself is continuous but its derivative has discontinuities. In other words, there are no piecewise analytic functions.

We suggest that the reader prove another consequence of Morera's theorem: *if a sequence of analytic functions defined in an open domain (G) is uniformly convergent* (the definition of uniform convergence is formulated for complex functions in the same way as for real functions; e.g. see [107], Sec. XVII.8) *the limit function of the sequence is also analytic in (G).*

6. Analytic Mappings and Maximum Modulus Principle. Let $w = f(z)$ be an analytic function in an open domain (G). As was said in Sec. 1.4, the corresponding mapping $z \to w$ (here and

henceforward a symbolic relation of this type designates a mapping from the z-plane into the w-plane) is termed analytic, and if $f'(z_0) \neq 0$ at a point $z_0 \in (G)$, the mapping of an infinitesimal neighbourhood of the point z_0 reduces, to within infinitesimals of higher order, to translation, uniform magnification and rotation. Now let us turn to the case when $f'(z_0) = 0$ but $f(z) \not\equiv \text{const}$. Then there is no linear term in Taylor's expansion in powers of $z - z_0$, and therefore we have

$$f(z) = f(z_0) + c_n(z - z_0)^n + \textit{terms of higher orders } (n \geqslant 2) \quad (17)$$

where the term $c_n(z - z_0)^n$ is the first one with a nonzero coefficient. Consequently, to within infinitesimals of higher order, we have $w - w_0 = c_n(z - z_0)^n$, and thus an infinitesimal neighbourhood of the point z_0 undergoes an n-valent mapping, i.e. the point w covers the corresponding infinitesimal neighbourhood of the point $w_0 = f(z_0)$ n times when the point z runs through the neighbourhood of the point z_0. A mapping of this type (namely, $w = z^k$) was described in Sec. 1.9. The additional multiplication by c_n determines a uniform magnification and rotation of the neighbourhood but the general geometric structure of the mapping is, in principle, the same. It is clear that in this case the mapping $z \longrightarrow w$ cannot be one-to-one.

What has been said implies a number of useful consequences. For example, it follows from formula (17) (which also holds for $n = 1$, i.e. for $f'(z_0) \neq 0$) that in a sufficiently small neighbourhood of the point z_0 the value $w_0 = f(z_0)$ is assumed by the function only at the point z_0, and $f(z) \neq w_0$ for the other points of the neighbourhood.

The last assertion, in its turn, implies the following property. Let a function $f(z)$ be analytic and not identically equal to zero in a finite closed region (K). By definition, this means that it is analytic in an open domain (G) containing (K). Then $f(z)$ *can only have a finite number of zeros in* (K). In fact, if otherwise, we can make use of the Bolzano-Weierstrass theorem asserting that from every infinite and bounded set of points of a Euclidean space it is possible to choose a convergent sequence of points (this theorem is proved in comprehensive courses of mathematical analysis meant for university students, but its assertion is intuitively clear). Thus we obtain an infinite sequence of zeros of $f(z)$ converging to a point belonging to (K) which is again a zero of $f(z)$ and thus, by the preceding paragraph, arrive at a contradiction (let the reader think about it!).

If a function $f(z)$ is analytic in the entire z-plane it may have an infinitude of zeros (for instance, this is the case for the function $f(z) = \sin z$). But, on the other hand, in every circle of finite radius it must have only a finite number of zeros. There-

fore, taking a sequence of circles, with infinitely increasing radii, which, in the limit, exhausts the whole plane, we conclude that *if the number of zeros of $f(z)$ is infinite, they can be arranged in a sequence $z_1, z_2, z_3, \ldots$, receding to infinity.*

Taking a difference of two functions and applying to it what has been proved we obtain the following assertion which is a generalization of Corollary 2 in Sec. 5: *if two functions are analytic in a domain (G) and their values coincide along an arc lying in (G), the functions coincide identically on (G).* In particular, this implies a simple proof of the possibility of passing from real relations to the corresponding complex relations (e.g. see [107], Sec. VIII.4). For instance, if we want to prove that $\sin^2 z = 1 - \cos^2 z$ we can reason as follows. The right-hand and left-hand members are analytic functions coinciding for real values of z, i.e. on the straight line $\operatorname{Im} z = 0$. Consequently, they coincide for all z, which is what we set out to prove.

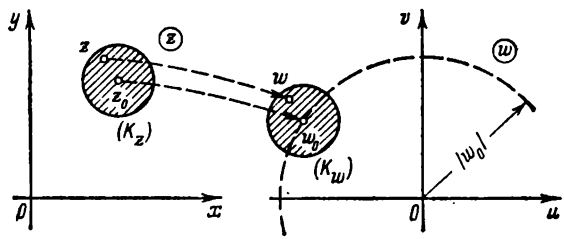


Fig. 34

Let a function $f(z)$ be analytic in a neighbourhood of a point z_0 and not identically equal to a constant. Then a small circle (K_z) with centre at z_0 is mapped, by the function $f(z)$ (to within infinitesimals of higher order; see Fig. 34), on the corresponding small circle K_w with centre at $w_0 = f(z_0)$. Let us choose in (K_w), in the vicinity of the point w_0, another point w for which $|w| > |w_0|$. The point w is the preimage of a point $z \in (K_z)$. Consequently, in every arbitrarily small neighbourhood of the point z_0 there is a point z such that $|f(z)| > |f(z_0)|$. Thus, the quantity $|f(z)|$ *cannot have a maximum at the point z_0* (let the reader carefully study this argument). This is the so-called **maximum modulus principle** of an analytic function.

Now we can draw the following consequence. Let $f(z)$ be an analytic function in a finite closed region (K). Then $|f(z)|$ is a real continuous function and hence it attains in (K) its greatest value (e.g. see [107], Sec. IX.4). It follows from the above that *this greatest value is necessarily attained on the boundary of (K).*

It can be similarly shown that $\operatorname{Re} f(z)$ cannot have z_0 as its point of extremum either (prove it!). It readily follows that a *function which is harmonic at all interior points of a finite closed domain and continuous everywhere including its boundary attains its maximum and minimum values on the boundary.*

Propositions of the type considered in the last two paragraphs are referred to as the *principles of the maximum.*

7. Analytic Continuation. Let a function $f(z)$ be analytic in a domain (G), and $g(z)$ in a domain (D), the domains (G) and (D) having a non-empty common portion (H) (Fig. 35) in which $f(z) \equiv g(z)$. Then the function

$$f_1(z) = \begin{cases} f(z) & \text{for} \quad z \in (G) \\ g(z) & \text{for} \quad z \in (D) \end{cases}$$

is obviously analytic in the domain (G_1) consisting of all the points of the domains (G) and (D). (By Sec. 5, even the condition that (G) and (D) contain a common line on which the limiting values of the functions $f(z)$ and $g(z)$ coincide and form a continuous function is sufficient for the function $f_1(z)$ to be analytic in the union of (G) and (D).) The function $f_1(z)$ is then called an **analytic continuation** of each of the functions $f(z)$ and $g(z)$. A continuation of this type is usually considered in the case when there are two formulas (determining analytic functions) which are applicable in two different domains having a common portion in which both formulas yield the same result.

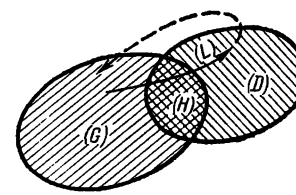


Fig. 35

The function $f_1(z)$ can in its turn be continued analytically to a wider region, etc. Theoretically, after all the possible continuations have been constructed we arrive at a **complete analytic function** $F(z)$ which cannot be continued to a wider domain. The function $F(z)$ is uniquely determined by the original function $f(z)$ since it readily follows from Sec. 6 that if the function $f(z)$ can be analytically continued along a curve (L), in the z-plane, which starts from (G), this can only be performed in a unique manner. But this curve (L), when extended, may again enter the domain (G) (Fig. 35), and then the new values of $f(z)$ in (G) do not necessarily coincide with the original values. In other words, if the original function is one-valued, its analytic continuation may be multiple-valued (and even infinite-valued).

An analytic continuation being always performed across an arc of the boundary of an original domain of definition, we can assert that a continued function cannot be extended to a wider region

if it is defined in the whole complex plane without cuts (although it can have singular points including branch points).

Let us consider some examples. The sum of the series

$$S(z)=1+z+z^2+\ldots+z^n+\ldots$$

which is convergent in the circle $|z|<1$ is an analytic function in that circle. But, for $|z|<1$ this sum is equal to

$$F(z)=\frac{1}{1-z}$$

the last expression representing a function which is analytic everywhere in the z-plane except the point $z=1$. Consequently, the complete analytic function $F(z)$ is an analytic continuation of the function $S(z)$. We say in these cases that $F(z)$ is an **extension** (continuation) of $S(z)$, and $S(z)$ is a **restriction** of $F(z)$.

Similarly, the infinite-valued function $\operatorname{Ln}(1+z)$ is a complete analytic function obtained as analytic continuation of the sum of the series $z-\frac{z^2}{2}+\frac{z^3}{3}-\ldots$, which is convergent in the same circle $|z|<1$.

It can be easily shown that the integral

$$\Gamma(p)=\int_0^\infty e^{-x}x^{p-1}dx=\int_0^\infty \exp\left[-x+(p-1)\ln x\right]dx \qquad (18)$$

determining the **gamma function** (e.g. see [107], Sec. XIV.17) is not only convergent for real values of p but also for imaginary p if $\operatorname{Re} p>0$. Since the rules for differentiating an integral with respect to a parameter (e.g. see [107], Sec. XIV.5) are readily extended to complex values of the parameter, integral (18) has a derivative with respect to p and therefore is an analytic function of p in the half-plane $\operatorname{Re} p>0$. As in the case of real values of p, it can readily be proved that in this half-plane we have $\Gamma(p+1)\equiv p\Gamma(p)$. Consequently, the function

$$\Gamma_1(p)=\frac{\Gamma(p+1)}{p}$$

which is defined in the half-plane $\operatorname{Re} p>-1$ and analytic there (why?) except the singular point $p=0$ coincides with $\Gamma(p)$ in the former half-plane. Thus, the function $\Gamma_1(p)$ is an analytic continuation of the function $\Gamma(p)$ from the former half-plane to the latter. Similarly, the function $\Gamma_2(p)=\frac{\Gamma_1(p+1)}{p}$ is an analytic continuation of the function $\Gamma_1(p)$ to the wider half-plane $\operatorname{Re} p>-2$. Proceeding in this manner we arrive at the complete gamma function which is also denoted by $\Gamma(p)$ and is a single-valued analytic function on the whole p-plane except the points

$p=0, -1, -2, \ldots$. The identity $\Gamma(p+1) \equiv p\Gamma(p)$ is fulfilled throughout the p-plane.

There is a general method of analytic continuation based on the symmetry principle (Sec. 1.17). This principle implies that if, for instance, the boundary of the domain of definition of an analytic function $f(z)$ contains a segment of the real axis, and the function $f(z)$ assumes real values on that segment, then the formula $f(z^*) = [f(z)]^*$ determines an analytic continuation of the function $f(z)$ across the segment.

If a function is given by its expansion into a power series, then the following **Euler transformation** can be of use for constructing an analytic continuation of the function. Suppose it is necessary to continue a function determined by an expansion of the type

$$f(z) = \gamma_0 c_0 + \gamma_1 c_1 z + \gamma_2 c_2 z^2 + \ldots \tag{19}$$

Consider another function $g(z)$ whose values in the entire z-plane are known and which has the expansion

$$g(z) = c_0 + c_1 z + c_2 z^2 + \ldots \tag{20}$$

at the point $z=0$ (i.e. $g(z)$ is regarded as being continued to the entire z-plane). In order to express $f(z)$ in terms of $g(z)$, let us multiply both sides of (20) by γ_0 and subtract the result from (19). This leads to

$$f - \gamma_0 g = (\gamma_1 - \gamma_0) c_1 z + (\gamma_2 - \gamma_0) c_2 z^2 + (\gamma_3 - \gamma_0) c_3 z^3 + \ldots \tag{21}$$

Now we differentiate both members of formula (20), multiply the result by $(\gamma_1 - \gamma_0) z = \Delta\gamma_0 z$ and subtract it from (21), which yields

$$f - \gamma_0 g - \Delta\gamma_0 z g' = (\gamma_2 - 2\gamma_1 + \gamma_0) c_2 z^2 + (\gamma_3 - 3\gamma_1 + 2\gamma_0) c_3 z^3 + \ldots \tag{22}$$

The next step is to differentiate (20) once again, multiply the result by $\frac{1}{2!}(\gamma_2 - 2\gamma_1 + \gamma_0) z^2 = \frac{1}{2!}\Delta^2\gamma_0 z^2$ and subtract it from (22), etc. It is readily proved that we can infinitely proceed in this way and thus, in the limiting process, arrive at the expansion

$$f(z) = \gamma_0 g(z) + \frac{\Delta\gamma_0}{1!} z g'(z) + \frac{\Delta^2\gamma_0}{2!} z^2 g''(z) + \ldots \tag{23}$$

which holds in a neighbourhood of the point $z=0$ if the absolute values of the coefficients γ_n do not grow too fast as $n \to \infty$ (the rate of growth must not be higher than that of an exponential expression). The right-hand side of (23) may turn out to be an analytic function defined in a domain which is wider than the circle of convergence of series (19), and then we obtain an analytic continuation of the function $f(z)$.

For instance, suppose that γ_n is a polynomial in n of degree k. Then $\Delta\gamma_n$ is a polynomial in n of degree $k-1$ (why?) and so

on. Therefore, the series on the right-hand side of (23) turns into a finite sum and hence we obtain a finite representation of the sum of series (19) in terms of the function $g(z)$, and thus the function $f(z)$ has been continued to the whole z-plane.

There is another special case that may be of use. Let us put

$$g(z)=(1+z)^{-p}$$

Then $c_0=1$, $c_1=-\frac{p}{1!}$, $c_2=\frac{p(p+1)}{2!}$, etc., and formulas (19) and (23) imply the relation

$$\gamma_0-\frac{p}{1!}\gamma_1 z+\frac{p(p+1)}{2!}\gamma_2 z^2-\ldots=$$

$$=(1+z)^{-p}\left[\gamma_0-\frac{p}{1!}\Delta\gamma_0\frac{z}{1+z}+\frac{p(p+1)}{2!}\Delta^2\gamma_0\left(\frac{z}{1+z}\right)^2-\ldots\right]$$

(check it up!). The result becomes particularly simple if we put $p=1$.

For example, let us take $p=1$ and $\gamma_n=\frac{1}{n+1}$. Then, computing, in succession, the differences, we readily derive the general expression $\Delta^k\gamma_n=\frac{(-1)^k k!}{(n+1)(n+2)\ldots(n+k+1)}$ whence $\Delta^k\gamma_0=\frac{(-1)^k}{k+1}$, that is

$$1-\frac{z}{2}+\frac{z^2}{3}-\frac{z^3}{4}+\ldots=\frac{1}{1+z}\left[1+\frac{1}{2}\frac{z}{z+1}+\frac{1}{3}\left(\frac{z}{z+1}\right)^2+\right.$$
$$\left.+\frac{1}{4}\left(\frac{z}{z+1}\right)^3+\ldots\right] \qquad (24)$$

The series on the left-hand side is convergent for $|z|<1$ $\left(\text{this is the expansion of the function } \frac{\ln(1+z)}{z} \text{ in powers of } z\right)$. The series on the right-hand side converges for $\left|\frac{z}{z+1}\right|<1$, i.e. for $|x+iy|^2<|x+iy+1|^2$, which yields $x>-\frac{1}{2}$. Hence, formula (24) gives an analytic continuation of the sum of the series on the left-hand side to the half-plane $\operatorname{Re} z>-\frac{1}{2}$.

8. Some Generalizations. *Analytic functions of a real variable.* If, according to the meaning of the definition, the argument x of a function $w=f(x)$ is a real variable the concept of analyticity introduced in Sec. 1.1. is inapplicable. In this case, based on results of Sec. 5, we say that *a function $f(x)$ is analytic in an interval $a<x<b$ if for every x_0 belonging to the interval it can be expanded into Taylor's series in powers of $x-x_0$ which is convergent in a neighbourhood of the point x_0.* To establish the relationship between this definition and the definition of Sec. 1.1 we

substitute the complex variable z for the real x entering into the Taylor expansion of the function $f(x)$ and thus obtain a function $f(z)$ analytic in a circle with centre at x_0 which assumes the given values $f(x)$ for $z=x$. Performing this operation for all values of x_0 we arrive at the function $f(z)$ defined in a domain of the complex z-plane containing the interval $a<x<b$ of the real axis (Fig. 36) which is analytic in this domain and coincides with the original function $f(x)$ for $z=x$. (When defining $f(z)$ in this manner for different x_0 we cannot arrive at a contradiction;

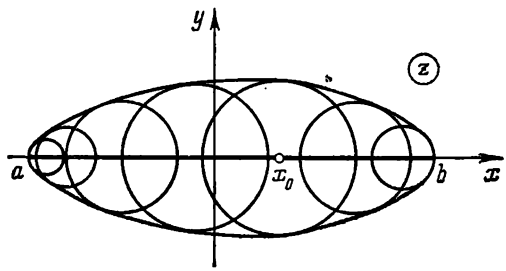


Fig. 36

let the reader prove this fact.) Thus, using the terminology of Sec. 7, we can say that an analytic function of a real variable can be extended (continued analytically) to an analytic function of a complex variable and, conversely, the former is a restriction of the latter.

Analytic functions of several variables. For simplicity, let us consider functions of two variables. *A function $f(z_1, z_2)$ of two complex variables z_1 and z_2 defined in an open domain (G) of the (complex) two-dimensional Cartesian space Z_2* (e.g. see [107], Sec. VII.20) *is said to be analytic in this domain if for every point $(z_{10}, z_{20}) \in (G)$ there is a neighbourhood in which the function can be expanded into a double Taylor series* (e.g. see [107], Sec. XVII.17) *in powers of (z_1-z_{10}) and (z_2-z_{20}).* The notion of an analytic function of two real arguments is introduced in like manner.

It should be noted that there is some kind of ambiguity in this terminology which sometimes may lead to misunderstanding. For instance, consider the function $w=y-x^2$ for real x and y. This is an analytic function which vanishes along the curve $y=x^2$ but is not identically equal to zero. One may think that this contradicts the properties of analytic functions established in Sec. 6. But in fact there is no contradiction here because this function is analytic as a function of real variables x and y (and even as a function of complex variables $x=z_1$ and $y=z_2$) but it is not an analytic function of the complex argument $z=x+iy$.

If $f(x_1, x_2)$ is an analytic function of real variables x_1 and x_2 it is also analytic as a function of each variable when the other is kept fixed. The converse is also true except some special cases which are practically unimportant. Moreover, if we have an analytic curve (L) with parametric equations $x_1 = \varphi_1(t)$, $x_2 = \varphi_2(t)$ (i. e. the functions φ_1 and φ_2 are analytic) which belongs to the domain of definition of an analytic function f, the composite function $f(\varphi_1(t), \varphi_2(t))$ is analytic as well. This property is implied by the possibility of substituting one power series into another (e. g. see [107], Sec. XVII.12). It follows that if a small arc of the curve (L) is a portion of a level line of the function f then the whole curve (L) lies in the level line (why?).

Implicit analytic functions are also of interest in some cases. For instance, let us consider an equation of the form

$$f(z_1, z_2) = 0 \tag{25}$$

where the function f is analytic. Suppose that the conditions

$$f(c_1, c_2) = 0, \quad f'_{z_2}(c_1, c_2) \neq 0 \tag{26}$$

are fulfilled (as is known, the second condition (26) is sufficient for existence of an implicit function; e. g. see [107], Sec. IX.13). Then it can easily be proved that in a sufficiently small neighbourhood of the point (c_1, c_2) equation (25) can be resolved in a unique manner with respect to z_2, and the function $z_2(z_1)$ thus obtained is also analytic, i. e. can be expanded into Taylor's series in powers of $z_1 - c_1$. The latter expansion can be found if we substitute $z_2 = z_2(z_1)$ into equality (25) and then differentiate, in succession, the resultant identity with respect to z_1 and compute the derivatives $\left.\frac{d^k z_2}{dz_1^k}\right|_{(c_1, c_2)}$ $(k = 1, 2, 3 \ldots)$. Another way to construct the expansion is to apply the method of undetermined coefficients. In particular, it follows that if the analytic function f is real, that is assumes real values, and the numbers c_1 and c_2 are also real, the analytic function $z_2(z_1)$ is real as well.

If the second condition (26) is violated the situation becomes more complicated but in the case of two independent variables it admits of a thorough investigation. For simplicity, let us suppose that $c_1 = c_2 = 0$ (if otherwise, we can introduce the new variables $\tilde{z}_k = z_k - c_k$ and thus reduce the latter case to the former) and consider the expansion of the function f in powers of z_1 and z_2. Some of the coefficients of the expansion can be equal to zero. Taking only the nonzero coefficients and numbering them in an arbitrary way we can write down this expansion in the form

$$f(z_1, z_2) = \sum_k d_k z_1^{m_k} z_2^{n_k} \tag{27}$$

where $d_k \neq 0$ and m_k and n_k are nonnegative integers ($k = 1, 2, \ldots$). It turns out that even in some simple cases (e. g. consider the function $f = z_1 + z_1^3 - z_2^3$) there can be no solution of equation (25) expressible as a series in integral powers of z_1. But at the same time this example (and other examples of that kind) indicates that it is advisable to try to construct a solution as a more general series

$$z_2 = \sum_{k=1}^{\infty} a_k s^k \quad (s^{q'} = z_1), \text{ i. e. } z_2 = \sum_{k=1}^{\infty} a_k z_1^{\frac{k}{q'}} \tag{28}$$

where q' is a positive integer (a series of type (28) is known as **Puiseux's series**). It can be proved (but we do not dwell on this question here) that series of form (28) represent the solutions of equation (25) in the most general case.

In practical applications it is sometimes sufficient to determine the dominant, principal, term in expansion (28) (i. e. the one with nonzero coefficient a_k for which the exponent k assumes the least value). This can be easily done because if $z_2 = As^p + \ldots$, $z_1 = s^q$ (where the dots designate the terms of higher order of smallness), relation (27) yields

$$f(z_1, z_2(z_1)) = \sum_k d_k A^{n_k} s^{pn_k + qm_k} + \ldots \tag{29}$$

Since the sum on the right-hand side of (29) is identically equal to zero (why?), all the summands must mutually cancel out. In particular, there must be at least two principal terms (i. e. with the lowest powers of s) in expression (29) since the principal terms cannot be eliminated by terms of higher order of smallness. But we have $d_k A^{n_k} \neq 0$, and therefore the integers $p \geqslant 1$ and $q \geqslant 1$ must be such that among the sums $pn_k + qm_k$ ($k = 1, 2, 3, \ldots$) there are at least two equal ones while the other sums are greater. If these values of p and q are known, the coefficient $a_p = A$ satisfies the algebraic equation

$$\sum_{pn_k + qm_k = \min} d_k A^{n_k} = 0 \tag{30}$$

(the sum in (30) is extended over all k for which the quantity $pn_k + qm_k$ attains its minimum value).

In concrete problems we can perform these calculations as follows. Let us consider the plane with coordinate axes m and n and mark on it the points (m_k, n_k) corresponding to all the combinations of the exponents in expansion (27). For instance, in Fig. 37 we see this construction for the function

$$f(z_1, z_2) \equiv z_1^6 - 2z_1^3 z_2^2 + 3z_1^2 z_2^3 + \frac{1}{2} z_1^5 z_2^3 - z_1 z_2^4 + z_1^2 z_2^5 - z_2^7 + z_2^8 \tag{31}$$

(check it up!). The first condition (26) indicates that the origin does not belong to the set of the marked points, that is there is no constant term in expansion (27). Furthermore, we can suppose that on each of the axes m and n there is at least one marked point because, if otherwise, equation (25) can be cancelled by the corresponding power of z_1 or z_2. If there is a straight line passing through at least two of the marked points which separates the other points from the origin we write down the equation of this line in the form $pn+qm=$ $=$const and thus obtain the sought-for values of p and q (which are taken as small as possible, that is without common factors). After p and q have been found we can determine the corresponding coefficient A from equation (30). (The segments of these lines form a polygonal line referred to as **Newton's polygon**; it is shown in heavy lines in Fig. 37.) Making this construction for equation (25) with function (31) (see Fig. 37) we conclude that the possible principal terms of the solutions are of the form

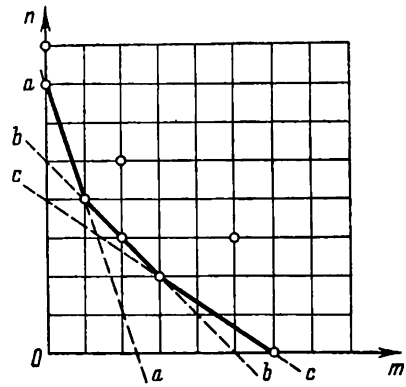


Fig. 37

$z_2=(\sqrt[3]{-1})_1 s$, $z_2=(\sqrt[3]{-1})_2 s$, $z_2=(\sqrt[3]{-1})_3 s$ where $s^3=z_1$ (for the straight line aa)

$z_2=2z_1$ and $z_2=z_1$ (for the straight line bb)

$z_2=\frac{1}{\sqrt{2}}s^3$, $z_2=\frac{-1}{\sqrt{2}}s^3$ where $s^2=z_1$ (for the straight line cc)

The first three formulas determine, in the case of complex z_1, the three-valued function $z_2=\sqrt[3]{-z_1}$. Similarly, the last two formulas can be rewritten as one formula $z_2=\sqrt{\frac{z_1^3}{2}}$. The fact that the number of solutions can be reduced by introducing multiple-valued functions always takes place when $q>1$ because, as is readily seen from Sec. 5, if a branch of a multiple-valued function satisfies an equation of type (25) the other branches satisfy it as well. By the way, if z_1 is regarded as a real variable (which often is the case in practical problems) then s should be understood as a concrete branch of the corresponding multiple-valued function, and

then all the seven solutions found above must be considered to be distinct.

Now we can proceed to determine subsequent terms of expansion (28) but it should be taken into account that the value of q' in formula (28) can either be equal to the value of q found for the principal term or have an additional integral factor which is unknown. Therefore we must perform the substitution $z_1 = s^q$, $z_2 = s^p(A + z_3)$ in equation (25) and thus pass to the corresponding equation for $z_3(s)$. If the conditions for the existence of an implicit function hold for the transformed equation, that is if in the corresponding Newton polygon the point (0, 1) is marked, the function z_3 is obtained in the form of a series in integral powers of s, which implies, in particular, that $q' = q$. But if these conditions do not hold we must again determine the possible principal terms of the expansion of $z_3(s)$, that is to put $z_3 = A_1 s_1^{p_1} + \ldots$, $s = s_1^{q_1}$, and then perform the substitution $s = s_1^{q_1}$, $z_3 = s_1^{p_1}(A_1 + z_4)$, etc. In practical problems, after a number of such substitutions, we usually arrive at the case when the second condition (26) is fulfilled, which enables us to apply a simpler technique (e. g. the method of undetermined coefficients) for further computations. In more comprehensive courses of complex analysis (e. g. see [17], [71], [133]) it is proved that the total number of solutions of equation (25) corresponding to one and the same principal term As^p is equal to the multiplicity of A as root of equation (30). In particular, it follows that the total number of solutions $z_2(z_1)$ of equation (25) in the vicinity of the origin is equal to the least exponent in the expansion of $f(0, z_2)$. If equations of type (30) only have simple roots for all principal terms of these solutions (e. g. this is the case in example (31)), the conditions for existence of implicit function are satisfied after the first of the above substitutions has been performed.

In applications we often deal with the case when all the variables concerned are real. Then the total number of solutions of equation (25) may be less than that indicated in the foregoing paragraph because a real equation may have several pairs of conjugate imaginary solutions (e. g. see [107], Sec. VIII.8) which should be rejected. For instance, in example (31) we have five solutions for $z_1 > 0$ and three for $z_1 < 0$. (It is clear that if the roots of equation (30) are simple and the leading coefficient is real then all the other coefficients together with the solutions are real as well.) If equation (25) involves a parameter, the number of real solutions may change in a jump-like manner when the parameter, in the process of its variation, takes on certain particular values. These values often play an important role in applied problems.

§ 3. SINGULAR POINTS AND ZEROS

1. Isolated Singular Points. In § 3 we shall consider only *isolated* singular points of single-valued analytic functions. A singular point is said to be **isolated** if there is a neighbourhood of the point in which the analyticity of the function is violated only at this point. The first paragraph of Sec. 2.5 implies that the function is necessarily unbounded in the vicinity of a point of this kind (although it does not necessarily approach infinity at that point; see. Sec. 2).

It should also be noted that there may exist so-called **removable singularities,** that is singular points which disappear after the function is redefined appropriately at these points (e. g. cf. [107], Sec. III.13). For example, the function $\frac{\sin z}{z}$ is not defined for $z=0$, that is from the formal point of view the point $z=0$ is a singularity of this function; but if we put $\left.\frac{\sin z}{z}\right|_{z=0}=1$ the singularity disappears since $\lim\limits_{z\to 0}\frac{\sin z}{z}=1$. In what follows we shall not regard such points as singular (i. e. we shall think of them as being removed in the above sense).

Let a single-valued function $f(z)$ have a singular point $z=a$. Then, according to Sec. 2.4, this function can be represented as the sum of Laurent's series in a neighbourhood of the point a:

$$f(z)=\sum_{k=-\infty}^{\infty} c_k(z-a)^k=\ldots+$$

$$+\frac{c_{-2}}{(z-a)^2}+\frac{c_{-1}}{z-a}+c_0+c_1(z-a)+c_2(z-a)^2+\ldots \qquad (1)$$

This series is convergent throughout the neighbourhood except the point a. The outer boundary of the maximal domain of convergence of the series is a circle with centre at a passing through another singular point of the function $f(z)$ nearest to a (if $f(z)$ is a one-valued branch of a multiple-valued function the branch points are also regarded as singular points). If there are no singular points other than a, series (1) converges throughout the complex z-plane for $z\neq a$.

It may turn out that the principal part of expansion (1) contains only a finite number of nonzero terms, that is there exists the first term in the expansion while there may not be the last one. A singular point of this kind is termed a **pole.** If there is an infinite number of terms with negative exponents we say that $z=a$ is an **essential singularity** of the function $f(z)$.

The coefficient c_{-1} in expansion (1) is called the **residue of the function $f(z)$ at its singular point $z=a$** and is denoted as $\operatorname{Res}_{z=a} f(z)$. The important role of this coefficient for computing integrals will be discussed in Sec. 2.3.

The expansion of a multiple-valued analytic function $f(z)$ in the vicinity of its singular point $z=a$ can be of a more complicated form. For instance, if we have a branch point of finite order n, we can introduce a new variable ζ determined by the relation $z-a=\zeta^n$ and find that if the point ζ describes a small circle with centre at $\zeta=0$ the point z makes n circuits about the point a, and therefore the function $f(z)$ returns to the original value. Hence, $f(z)$ is a single-valued analytic function of ζ and can be expanded into Laurent's series

$$f(z)=\sum_{k=-\infty}^{\infty} tc_k{}^k=\sum_{k=-\infty}^{\infty} c_k (z-a)^{\frac{k}{n}} \tag{2}$$

The series $\sum_{k=-\infty}^{\infty} c_k(z-a)^{\frac{k}{n}}$ in integral powers of the quantity $(z-a)^{\frac{1}{n}}$ is sometimes referred to as the **Laurent-Puiseux** series. Making $(z-a)^{\frac{1}{n}}$ assume all its n values we obtain all the values of the function $f(z)$.

2. Pole. The expansion of a function into Laurent's series in the vicinity of its pole is of the form

$$f(z)=\frac{c_{-n}}{(z-a)^n}+\frac{c_{-n+1}}{(z-a)^{n-1}}+\ldots+\frac{c_{-1}}{z-a}+c_0+c_1(z-a)+\ldots \tag{3}$$

where $c_{-n}\neq 0$. We then say that the pole is of **order** n. If $n=1$ the pole is said to be **simple**, if otherwise (i.e. of order 2, 3, etc.) it is called **multiple.** For a pole of order n the term $\frac{c_{-n}}{(z-a)^n}$ in expansion (3) is dominant when $z\to a$ (why?). For $z\to a$ the quantity $f(z)$ is infinitely large and equivalent to $\frac{c_{-n}}{(z-a)^n}$, that is of order n relative to $\frac{1}{z-a}$ (e.g. cf. [107], Sec. III.11). It follows, in particular, that every function must approach infinity at its pole.

Poles often appear when we consider a quotient

$$f(z)=\frac{g(z)}{h(z)} \tag{4}$$

of two analytic functions for which a is a **regular** (non-singular) point, and $h(a)=0$. If in this case we have $g(a)\neq 0$, the

point $z=a$ is a pole of the function $f(z)$, its order being equal to the multiplicity of a as root of $h(z)$ (see Sec. 2.5). If $g(a)\neq 0$, and $z=a$ is a zero of order p for the function $g(z)$ and of order $q>p$ for the function $h(z)$, the function $f(z)$ has a pole of order $q-p$ at the point $z=a$ (what happens in the case $q\leqslant p$?).

These assertions are readily proved if we expand $g(z)$ and $h(z)$ in powers of $z-a$, reduce fraction (4) by the greatest possible power of $z-a$ and then perform the division of the series (e.g. see [107], Sec. XVII.12).

It is easily proved that if a single-valued analytic function (which originally is not necessarily given in form (4)) has a finite order of growth relative to $\frac{1}{z-a}$ for $z\to a$, that is if $f(z)=$ $=O\left(\frac{1}{(z-a)^n}\right)$, the function $g(z)=(z-a)^n f(z)$ no longer has a singularity at the point $z=a$ and therefore this point is a pole of the function $f(z)$.

The residue of a function at a pole can be determined in a particularly simple way when the pole is simple. Indeed, if

$$f(z)=\frac{c_{-1}}{z-a}+c_0+c_1(z-a)+\ldots$$

then $(z-a)f(z)=c_{-1}+c_0(z-a)+c_1(z-a)^2+\ldots$, and therefore

$$\operatorname*{Res}_{z=a} f(z)=c_{-1}=\lim_{z\to a}[(z-a)f(z)]$$

Taking the case (4), we see that if $g(a)\neq 0$, $h(a)=0$ and $h'(a)\neq 0$ (the latter condition means that the denominator has a zero of order one), we can write

$$\operatorname*{Res}_{z=a} f(z)=\lim_{z\to a}\frac{g(z)(z-a)}{h(z)}=\lim_{z\to a}\frac{g'(z)(z-a)+g(z)}{h'(z)}=\frac{g(a)}{h'(a)} \qquad (5)$$

Here we have had an indeterminate form of type $\frac{0}{0}$ and have used L'Hospital's rule which applies to analytic functions of a complex variable as well (the proof of the rule is based on Taylor's formula, and we leave it to the reader).

For the general case of a pole of order n we obtain from (3) the relation

$$(z-a)^n f(z)=c_{-n}+c_{-n+1}(z-a)+\ldots+c_{-1}(z-a)^{n-1}+c_0(z-a)^n+\ldots$$

Differentiating both sides $n-1$ times we arrive at the equality

$$[(z-a)^n f(z)]^{(n-1)}=(n-1)!c_{-1}+\textit{the terms with positive powers of } z-a$$

(check up the calculations!). Now, passing to the limit for $z\to a$ we derive the desired formula

$$\operatorname*{Res}_{z=a} f(z)=c_{-1}=\frac{1}{(n-1)!}\lim_{z\to a}[(z-a)^n f(z)]^{(n-1)} \qquad (6)$$

When applying this formula we must determine the order n of the pole as was described above. Formula (6) remains valid for a pole of order less than n (but not higher than n) because the quantity c_{-n} in the preceding argument may be equal to zero.

The behaviour of a function $f(z)$ in the vicinity of its essential singularity $z=a$ is much more complicated. Formula (1) shows that in this case the values of $f(z)$ cannot be bounded for $z \to a$ by a power function but it turns out that nevertheless they are not infinitely large. Moreover, it can be easily proved that the function $\frac{1}{f(z)-k}$, where k is any complex number, cannot be bounded in the vicinity of the point a (if we suppose that it is bounded we can take its Taylor's expansion and thus arrive at a contradiction; let the reader carefully examine this method of proving the assertion). But this means that near the point a the function $f(z)$ assumes values which are arbitrarily close to any number. This interesting theorem was proved in 1868 by Yu. V. Sokhotsky (1842-1929), a Russian mathematician. This theorem shows that it would be incorrect to think of an essential singularity as a pole of infinite order!

As an example of an essential singularity we can take the point $z=0$ for the function

$$e^{\frac{1}{z}} = 1 + \frac{1}{z} + \frac{1}{2!z^2} + \frac{1}{3!z^3} + \dots \tag{7}$$

Let the reader consider the character of variation of the function for various ways in which the point z may approach the origin.

Formulas of type (6) no longer hold for essential singularities. By the way, if we know the Laurent expansion of a function in the vicinity of its essential singularity the residue of the function at that point is readily determined. For instance, formula (7) indicates that $\operatorname{Res}_{z=0} e^{\frac{1}{z}} = 1$.

3. Cauchy's Residue Theorem. One of the important applications of residues is the computation of integrals of the form

$$\oint_{(L)} f(z)\,dz$$

taken over closed contours. Let $f(z)$ be a single-valued analytic function everywhere on (L) and inside (L) except a finite number of singular points $z_1, z_2, \dots, z_N$ lying in the interior of (L), the contour (L) being positively oriented. Let us break up the region bounded by the contour (and lying inside it) into N parts in such a way that each part contains only one singular point (see Fig. 38 where the case $N=3$ is depicted). Considering the contours

(L_k) $(k=1, 2, \ldots, N)$ bounding these parts to be positively oriented we can write the relation

$$\oint_{(L)} f(z)\,dz = \sum_{k=1}^{N} \oint_{(L_k)} f(z)\,dz \tag{8}$$

because the integrals on the right-hand side taken along the arcs shown in dotted line mutually cancel out (why?). Now we can transform each contour (L_k) into a small circle with centre at z_k (in Fig. 38 a deformation of this type is shown for $k=1$) without affecting the values of the integrals (see Sec. 2.2). For a sufficiently small circle Laurent's expansion of $f(z)$ at z_k is valid, and therefore, by (2.10), we obtain, for this Laurent's series, the relation

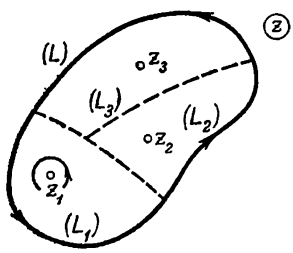


Fig. 38

$$\oint_{(L_k)} f(z)\,dz = 2\pi i \cdot c_{-1} = 2\pi i \cdot \operatorname*{Res}_{z=z_k} f(z)$$

Taking these expressions for all $k=1, 2, \ldots, N$ and substituting them into (8) we arrive at **Cauchy's residue theorem**:

$$\oint_{(L)} f(z)\,dz = 2\pi i \sum_{k=1}^{N} \operatorname*{Res}_{z=z_k} f(z) \tag{9}$$

It should be noted that on the right-hand side of (9) the summation is extended over all singular points of the function $f(z)$ lying inside (L). If there are no such singular points the integral is equal to zero according to Cauchy's integral theorem (given in Sec. 2.2). Thus, the latter can be regarded as a special case of the general theorem (9) if we agree that, by definition, *the residue of an analytic function at its regular point is equal to zero.* This convention is also convenient when we do not know beforehand whether a point is singular or when there is a removable singularity (which we agreed to regard as a regular point after the function is redefined appropriately). Under this convention formulas (5) and (6) remain valid.

Cauchy's residue theorem makes it possible to determine the exact values of integrals without finding the corresponding indefinite integrals which may have complicated expressions or may be irreducible to elementary functions. Therefore, the application of this theorem is extremely useful in various divisions of mathematical analysis. It should be noted that the theorem can also be applied to computing real integrals by reducing them in an artificial manner to complex ones.

Let us consider a simple example. Suppose that it is necessary to compute the integral

$$I = \int_0^{2\pi} \frac{dt}{a + b\cos t + c\sin t} \tag{10}$$

on condition that $b^2 + c^2 < a^2$ which is equivalent to the requirement that the denominator of the integrand should not vanish. Applying Euler's formulas we can pass from the trigonometric functions to the corresponding exponential expressions and make the substitution $e^{it} = z$ which yields (check it up!)

$$I = \oint_{(L)} \frac{2}{(bi + c)z^2 + 2iaz + (bi - c)} dz$$

where (L) is a unit circle oriented in the positive direction. The integrand in the last integral has two simple poles, only one of them, determined by the formula

$$z_1 = -\frac{a - \sqrt{a^2 - (b^2 + c^2)}}{bi + c} i$$

being inside (L). Now applying formulas (5) and (9) we finally obtain

$$I = \frac{2\pi}{\sqrt{a^2 - (b^2 + c^2)}}$$

In this example we have essentially used the fact that integral (10) is taken over an interval whose length is equal to the period of the integrand (this makes it possible to reduce the original integral to one taken over a closed contour). For other limits of integration a contour of type (L) becomes nonclosed and then formula (9) is inapplicable.

Another consequence of Cauchy's residue theorem is **Cauchy's integral formula** which expresses the values of an analytic function inside a closed contour in terms of its boundary values assumed on the contour. To derive this formula let us consider the integral

$$\oint_{(L)} \frac{f(z)}{z - z_0} dz \tag{11}$$

where the function $f(z)$ is one-valued and analytic everywhere inside the contour (L) and on it, and the point z_0 is taken inside (L). Here the integrand function $\frac{f(z)}{z - z_0}$ has exactly one singular point inside the contour (L), namely the simple pole $z = z_0$, the residue at z_0 being equal to $f(z_0)$ (why?). Consequently, by Cauchy's

residue theorem, integral (11) is equal to $2\pi i f(z_0)$ whence we obtain Cauchy's integral formula

$$f(z_0)=\frac{1}{2\pi i}\oint\limits_{(L)}\frac{f(z)}{z-z_0}dz \tag{12}$$

(Find the value of the right-hand member in the case when the point z_0 is outside (L).)

Differentiating both sides of (12) with respect to the parameter z_0 we obtain the formulas

$$f'(z_0)=\frac{1!}{2\pi i}\oint\limits_{(L)}\frac{f(z)}{(z-z_0)^2}dz, \quad f''(z_0)=\frac{2!}{2\pi i}\oint\limits_{(L)}\frac{f(z)}{(z-z_0)^3}dz, \quad \ldots$$

which are sometimes used for estimating the derivatives.

4. Application to Improper Integrals. Cauchy's residue theorem is widely applied to computing improper integrals, which sometimes involves sophisticated analytical methods. Here we shall only indicate some simple results related to these problems. For more detail we refer the reader to the general courses indicated above (in particular, see [81]) and special monographs.

Let us consider a convergent improper integral of the general form

$$I=\int\limits_{-\infty}^{\infty}f(x)\,dx \tag{13}$$

under the assumption that the function f is not only defined for real but also for imaginary values of the argument and that $f(z)$ is a single-valued analytic function in the half-plane $\operatorname{Im} z\geqslant 0$ except a finite number of singular points $z_1, z_2, \ldots, z_N$, all the imaginary parts $\operatorname{Im} z_k$ being positive. The original problem does not involve any closed contour and therefore the residue theorem cannot be directly applied to integral (13). Therefore, we introduce an auxiliary sequence of arcs $(L_1), (L_2), (L_3), \ldots$ (see Fig. 39) such that the corresponding contours $(L_n)'$ formed of the segments $[a_n, b_n]$ of the x-axis and the arcs (L_n) are closed. Suppose that the arcs (L_n) are chosen in such a way that for the given function $f(z)$ the following conditions are fulfilled:

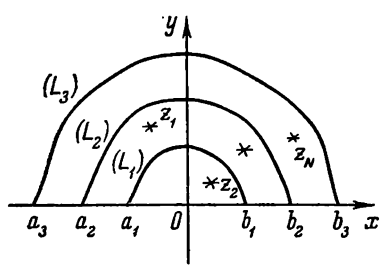


Fig. 39

1. $a_n\to-\infty$ and $b_n\to\infty$ for $n\to\infty$.

2. The singular points $z_1, z_2, \ldots, z_N$ are inside $(L_n)'$ for sufficiently large n.

3. $\int\limits_{(L_n)} f(z)\,dz \to 0$ for $n \to \infty$.

Then integral (13) can be easily computed. Indeed, by the residue theorem and Condition 2, we can write, for sufficiently large n, the relation

$$\int\limits_{(L_n)'} f(z)\,dz = \int\limits_{a_n}^{b_n} f(x)\,dx + \int\limits_{(L_n)} f(z)\,dz = 2\pi i \sum_{k=1}^{N} \operatorname*{Res}_{z=z_k} f(z) \qquad (14)$$

Now making n tend to infinity we see that according to Condition 1 the term $\int\limits_{a_n}^{b_n} f(x)\,dx$ tends to I while, by Condition 3, the term $\int\limits_{(L_n)} f(z)\,dz$ tends to zero. Thus, in the limit we obtain the expression

$$\int\limits_{-\infty}^{\infty} f(x)\,dx = 2\pi i \sum_{k=1}^{N} \operatorname*{Res}_{z=z_k} f(z) \qquad (15)$$

The sequence of arcs (L_n) with the desired properties is readily constructed for some classes of integrals (13). For example, suppose that when z approaches infinity in the upper half-plane, the function $f(z)$ tends to zero faster than $\frac{1}{z}$, that is

$$f(z) = o\left(\frac{1}{z}\right) \qquad (z \to \infty,\ \operatorname{Im} z \geqslant 0) \qquad (16)$$

(e.g. see [107], Sec. III.11 p. 84 on this notation). Then we can simply take, as contours (L_n), the semicircular arcs of radii n $(n = 1, 2, 3, \ldots)$ with centres at the origin. In fact, for these semicircles Conditions 1 and 2 are obviously fulfilled and Condition 3 is implied by relation (2.3).

Example. Let us compute the integral

$$I = \int\limits_{-\infty}^{\infty} \frac{1}{x^2+p^2} e^{i\omega x}\,dx \qquad (17)$$

where $\omega \geqslant 0$ and $p > 0$ are real. This integral is often applied in physics. Note that the corresponding indefinite integral is inexpressible in terms of elementary functions. Putting $z = x + iy$, $y \geqslant 0$, we can write

$$\left|\frac{1}{z^2+p^2}\, e^{i\omega z}\right| = \left|\frac{1}{z^2+p^2}\right| |e^{i\omega x}|\,|e^{-\omega y}| \leqslant \left|\frac{1}{z^2+p^2}\right| = O\left(\frac{1}{z^2}\right) \ (z \to \infty)$$

and hence condition (16) is fulfilled. Consequently, by formula (15), we have

$$I = 2\pi i \operatorname*{Res}_{z=pi} \frac{e^{i\omega z}}{z^2+p^2} = 2\pi i \cdot \frac{e^{i\omega pi}}{2pi} = \frac{\pi}{p} e^{-\omega p}$$

If the function $f(z)$ has a number of real isolated singular points $x_1, x_2, \ldots x_M$ integral (13) is **singular** (e.g. see [107], Sec. XIV.19). It is readily proved that if the principal parts of Laurent's expansions of the function $f(z)$ at these points only involve odd powers, integral (13) possesses *Cauchy's principal value*, and, instead of formula (15), we have

$$\text{v.p.} \int_{-\infty}^{\infty} f(x)\,dx = 2\pi i \sum_{k=1}^{N} \operatorname*{Res}_{z=z_k} f(z) + \pi i \sum_{k=1}^{M} \operatorname*{Res}_{z=x_k} f(z) \qquad (18)$$

Indeed, let, for simplicity, $M = 1$. Choose, instead of the contours shown in Fig. 39, the contours depicted in Fig. 40 which contain additional semicircular arcs (γ_n) of radius ε_n; let ε_n tend to zero in an arbitrary manner when $n \to \infty$. Then we have, in formula (14), the expression

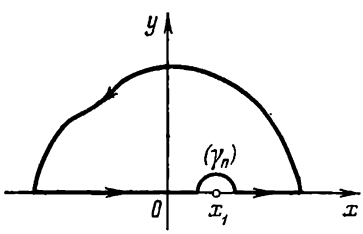


Fig. 40

$$\int_{a_n}^{x_1-\varepsilon_n} f(x)\,dx + \int_{(\gamma_n)} f(z)\,dz + \int_{x_1+\varepsilon_n}^{b_n} f(x)\,dx$$

instead of the term $\int_{a_n}^{b_n} f(x)\,dx$. But, according to the hypothesis, the second integral in this expression tends to $-\pi i$ for $n \to \infty$ (why?), which implies formula (18). We see that the singular points x_i lying on the path of integration yield the terms of the type $\pi i \operatorname*{Res}_{z=x_i} f(z)$ instead of the values $2\pi i \operatorname*{Res}_{z=z_k} f(z)$ which correspond to the singular points z_k located above the x-axis. Let the reader show that if x_1 is a pole, and the principal part of Laurent's expansion at this point contains at least one even power, integral (13) does not possess Cauchy's principal value.

Now let us generalize integral (17) and consider integrals of the type

$$\int_{-\infty}^{\infty} f(x)\,e^{i\omega x}\,dx \quad (\omega > 0) \qquad (19)$$

It turns out that for the above method to be applicable it is sufficient to introduce the condition

$$f(z) \longrightarrow 0 \qquad (z \longrightarrow \infty,\ \operatorname{Im} z \geqslant 0) \tag{20}$$

which is less restrictive than the former requirement (16). In particular, it follows that any proper rational fraction without real poles can be taken as the function $f(x)$; in this case the only difficulty connected with the computation of integral (19) is to determine the zeros of the denominator of the fraction.

To prove the applicability of the method we must show that if condition (20) is fulfilled, the corresponding integral

$$I_R = \int\limits_{(L_R)} f(z)\, e^{i\omega z}\, dz$$

taken over the "upper" semicircular arc of radius R with centre at the point $z=0$ tends to zero for $R \longrightarrow \infty$. To this end, we put $z = Re^{i\varphi}$ and apply the first inequality (2.3):

$$|I_R| \leqslant \int\limits_0^{\pi} |f(Re^{i\varphi})|\, |\exp(i\omega R\cos\varphi - \omega R\sin\varphi)|\, R\, d\varphi =$$

$$= R\int\limits_0^{\pi} |f(Re^{i\varphi})|\exp(-\omega R\sin\varphi)\, d\varphi \leqslant$$

$$\leqslant \max_{(L_R)} |f(z)| \cdot R\int\limits_0^{\pi} \exp(-\omega R\sin\varphi)\, d\varphi$$

By condition (20), the first factor on the right-hand side of the latter expression tends to zero for $R \longrightarrow \infty$, and therefore it is sufficient to show that the second factor is bounded. We have

$$R\int\limits_0^{\pi} \exp(-\omega R\sin\varphi)\, d\varphi = 2R\int\limits_0^{\pi/2} \exp\left(-\omega R\frac{\sin\varphi}{\varphi}\varphi\right) d\varphi \leqslant$$

$$\leqslant 2R\int\limits_0^{\pi/2} \exp\left(-\omega R\frac{2}{\pi}\varphi\right) d\varphi \tag{21}$$

(here we have used the inequality

$$\frac{\sin\varphi}{\varphi} > \frac{\sin\frac{\pi}{2}}{\left(\frac{\pi}{2}\right)} = \frac{2}{\pi} \qquad \left(0 < \varphi < \frac{\pi}{2}\right)$$

which can easily be established by considering the graph of the sine). We see that the right-hand side of (21) is less than

$$2R\int\limits_0^{\infty} \exp\left(-\omega R\frac{2}{\pi}\varphi\right) d\varphi = \frac{\pi}{\omega} = \text{const}$$

which implies the validity of the condition $I_R \xrightarrow[R \to \infty]{} 0$ provided that condition (20) is fulfilled. Thus, the proposition asserting that condition (20) implies $I_R \xrightarrow[R \to \infty]{} 0$ has been proved. It is called **Jordan's lemma** after C. Jordan (1838-1922), a noted French mathematician. From this assertion we can easily derive the relation

$$\int\limits_{(M_R)} f(z)\, e^{\omega z}\, dz \xrightarrow[R \to \infty]{} 0 \qquad (\omega > 0) \tag{22}$$

on condition that $f(z) \to 0$ for $z \to \infty$, $\operatorname{Re} z \leqslant 0$, where (M_R) is the "left" semicircular arc of radius R with centre at the point $z = 0$. The latter assertion is also referred to as Jordan's lemma.

If condition (20) is fulfilled, integral (19) is convergent (but in the general case, not absolutely convergent!) which can be easily proved (e. g. see [107], Sec. XIV.15 where a similar proof is given for the integral $\int\limits_1^\infty \frac{1}{x} \sin x\, dx$).

If the function $f(z)$ has, in the upper half-plane (i.e. for $\operatorname{Im} z > 0$), an infinite sequence of singular points tending to infinity, then, to apply the above method, we must make the radius R run over a specifically chosen sequence $R_1, R_2, \ldots, R_n, \ldots$ ($\lim\limits_{n \to \infty} R_n = \infty$) since in this case the quantity R cannot vary in an arbitrary way because the corresponding semicircular arcs must not pass through the singular points. The above proof indicates that it is sufficient to require that relation (20) hold only for the semicircles (L_{R_n}), i. e. we can take the condition

$$\max_{0 \leqslant \varphi \leqslant \pi} |f(R_n e^{i\varphi})| \xrightarrow[n \to \infty]{} 0$$

because the behaviour of $f(z)$ between these semicircular arcs is inessential. Applying the residue theorem and passing to the limit for $n \to \infty$ we arrive at the final formula for integral (19) which involves the sum of an infinite series. What has been said also applies to formula (15).

When this method is applied to an integral dependent on a parameter it is sometimes necessary to introduce different closed contours for different values of the parameter. For example, in physics we encounter integrals of the form

$$I = \int\limits_{(L)} f(z)\, e^{i\omega z}\, dz \tag{23}$$

where the function $f(z)$ is single-valued and analytic in the whole

z-plane, except a finite number of singular points some of which may be real (see Fig. 41 where the singular points are marked by small circles), and satisfies the condition $f(z) \xrightarrow[z \to \infty]{} 0$. The contour (L) (Fig. 41) coincides with the real axis everywhere except small neighbourhoods of the singular points located on the x-axis where it includes small arcs drawn in such a way that these singular points lie above or below them. Suppose that $\omega \neq 0$ is a real number. Then, if $\omega > 0$, we form a closed contour by adding to (L) an upper semicircle of large radius and apply Jordan's lemma, which results in the expression

$$I = 2\pi i \sum_{\text{above } z=z_s} \operatorname{Res}\left[f(z)\, e^{i\omega z}\right] \qquad (\omega > 0)$$

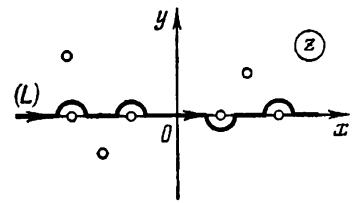


Fig. 41

where the sum is extended over all singular points z_s of the function $f(z)$ lying above (L) (including those lying on the x-axis). If $\omega < 0$ we take the lower semicircle of large radius and integrate over the closed contour (L) thus obtained in the negative direction. Then after the passage to the limit has been performed we obtain the formula

$$I = -2\pi i \sum_{\text{below } z=z_s} \operatorname{Res}\left[f(z)\, e^{i\omega z}\right] \qquad (\omega < 0)$$

where the sum is extended over all singular points lying below (L).

If we deal with a multiple-valued function (which may appear when the integrand is analytically continued) we must carefully distinguish between its branches. As an example, let us consider the real integral

$$\int_0^\infty f(x)\, x^{a-1}\, dx \qquad (0 < a < 1) \tag{24}$$

where the function $f(x)$ can be analytically extended to the entire z-plane $(z = x + iy)$ in such a way that the continued function $f(z)$ is single-valued and analytic everywhere except its singular points $z_1, z_2, \ldots, z_N$ which do not lie on the positive half-axis $0 \leqslant x < \infty$. In order to compute this integral we introduce the auxiliary complex integral

$$I_R = \int_{(S_{Rr})} f(z)\, z^{a-1}\, dz \tag{25}$$

where the closed contour (S_{Rr}) (shown in Fig. 42) consists of two circles of radii R and r, respectively, and the line segment $r \leqslant x \leqslant R$ which is traversed twice (in the opposite directions;

in Fig. 42 we have two "replicas" of this segment which are shown so that the reader could visualize the contour). The integrand in (25) being a multiple-valued function, we must stipulate what branch of the function is meant. For this purpose we make a cut along the semiaxis $0 \leqslant x < \infty$ and, for definiteness, choose the branch of the function z^{a-1} which assumes the value x^{a-1} on the upper edge of the slit. It should be noted (this is very important!) that the contour S_{Rr} does not intersect the cut. On the lower edge of the slit the branch we have chosen assumes the values $e^{2\pi i (a-1)} x^{a-1}$ (why?), and therefore, in contrast to the case of a single-valued integrand, the two integrals along the line segment $r \leqslant x \leqslant R$ taken in the opposite directions do not mutually cancel out. This fact enables us to compute integral (24).

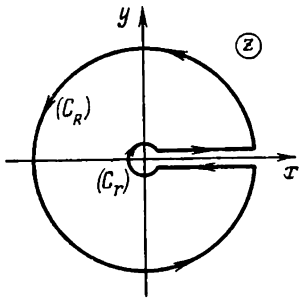


Fig. 42

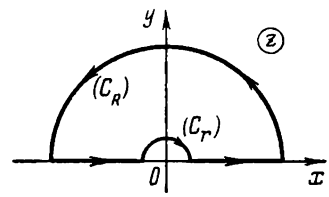


Fig. 43

For a sufficiently small r and a sufficiently large R we have

$$I_R = \int\limits_{(C_r)} f(z)\, z^{a-1}\, dz + \int\limits_{(C_R)} f(z)\, z^{a-1}\, dz + (1 - e^{2\pi i (a-1)}) \int\limits_r^R f(x)\, x^{a-1}\, dx =$$

$$= 2\pi i \sum_{k=1}^{N} \operatorname*{Res}_{z=z_k} [f(z)\, z^{a-1}] \qquad (26)$$

But $\lim\limits_{r \to 0} \int\limits_{(C_r)} f(z)\, z^{a-1}\, dz = 0$ (why?), and therefore, if

$$\int\limits_{(C_R)} f(z)\, z^{a-1}\, dz \xrightarrow[R \to \infty]{} 0$$

we can pass to the limit for $r \to 0$, $R \to \infty$ and thus conclude that integral (24) is convergent and equal to the right-hand side of (26) divided by $1 - e^{2\pi i (a-1)}$.

If, in addition, we suppose that $f(x)$ is an even or an odd function, that is $f(-x) \equiv \varepsilon f(x)$ where $\varepsilon = 1$ or $\varepsilon = -1$, we can use the contour shown in Fig. 43, which yields an analogous result with the divisor $1 + \varepsilon e^{\pi i (a-1)}$ (instead of $1 - e^{2\pi i (a-1)}$) and the

sum of residues extended over the singular points in the upper half-plane.

In particular, applying the above result to the integral $\int_0^\infty \frac{x^{a-1}}{1+x}dx$ for $0<a<1$ we obtain (check it up!)

$$\int_0^\infty \frac{x^{a-1}}{1+x}dx = \frac{2\pi i}{1-e^{2\pi i(a-1)}} \operatorname*{Res}_{z=-1}\left(\frac{z^{a-1}}{1+z}\right) = \frac{2\pi i}{1-e^{2\pi i(a-1)}} e^{\pi i(a-1)} = \frac{\pi}{\sin \pi a}$$

But, as is known (e. g. see [107], formulas (XIV.70) and (XIV.73)), this integral is equal to $B(a, 1-a) = \Gamma(a)\Gamma(1-a)$. Equating these results and making use of the theorem in Sec. 2.6 on the coincidence of two analytic functions assuming the same values along a line we arrive at the important identity

$$\Gamma(z)\Gamma(1-z) \equiv \frac{\pi}{\sin \pi z} \tag{27}$$

for the gamma function which holds for any complex z. In particular, it follows that the function $\Gamma(z)$ has no zeros, and the points $z=-n$ $(n=0, 1, 2, \ldots)$ are its simple poles with residues $\frac{(-1)^n}{n!}$ (why?).

By analogy with integral (24), we can compute the integral

$$\int_{-1}^{1}\left(\frac{1-x}{1+x}\right)^\alpha f(x)\,dx \qquad (-1<\alpha<1,\ \alpha\neq 0)$$

where the function $f(z)$ is one-valued and analytic throughout the z-plane except a finite number of singular points not lying on the path of integration and satisfies condition (16) at infinity. For this purpose it is advisable to take the contour of integration consisting of two ellipses, with focuses at the points $z=\pm 1$, whose eccentricities tend, respectively, to 0 and to 1. Let the reader prove that this integral is equal to

$$\frac{\pi}{\sin \pi\alpha}\sum \operatorname{Res}\left[\left(\frac{z-1}{z+1}\right)^\alpha f(z)\right] \tag{28}$$

where the sum is extended over all singular points of the function $f(z)$, and the expression $\left(\frac{z-1}{z+1}\right)^\alpha$ is understood as the branch which is one-valued outside the interval of integration and is equal to unity for $z=\infty$. (If condition (16) is violated the sum on the right-hand side of (28) must additionally include the residue at $z=\infty$; see Sec. 6.)

There are many other types of integral that can be computed with the aid of the residue theorem for appropriately chosen contours of integration.

The computation of the sum of the series

$$S=\sum_{k=-\infty}^{\infty} f(k)$$

where $f(z)$ is a single-valued analytic function in the whole z-plane except its isolated singular points $z_j \neq 0, \pm 1, \pm 2, \ldots$ is closely related to the problem of computing improper integrals (some additional conditions on $f(z)$ will be enumerated later on). Consider the auxiliary integral

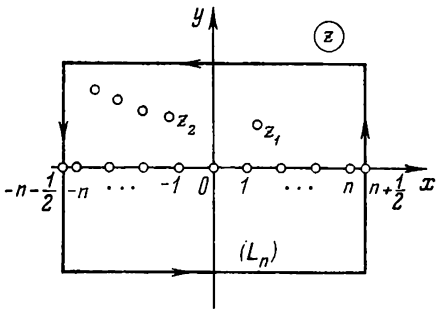


Fig. 44

$$I_n=\int_{(L_n)} \cot(\pi z)\, f(z)\, dz \qquad (29)$$

taken along the contour (L_n) shown in Fig. 44. We shall suppose that there are no singular points of the function $f(z)$ on (L_n) and that

$$\int_{(L_n)} |f(z)|\,|dz| \xrightarrow[n\to\infty]{} 0$$

Then we have $I_n \xrightarrow[n\to\infty]{} 0$ because the function $\cot \pi z$ is uniformly bounded on these contours. But, according to Cauchy's residue theorem, the integral I_n is equal to the expression

$$2\pi i\left[\sum_{k=-n}^{n} \frac{1}{n} f(n)+\sum_{j} \operatorname*{Res}_{z=z_j} \cot(\pi z)\, f(z)\right]$$

where the second sum is extended over all singular points of the function $f(z)$ lying inside (L_n). Passing to the limit for $n \to \infty$ we derive the formula

$$\sum_{k=-\infty}^{\infty} f(n)=-\pi \sum_{j} \operatorname*{Res}_{z=z_j} \cot(\pi z)\, f(z) \qquad (30)$$

where the sum on the right-hand side extends over all the singular points of the function $f(z)$ in the z-plane. If there is an infinitude of singular points we should additionally require that the series on the right-hand side be convergent.

If the function $f(z)$ has singular points $z=n$ where the numbers n are integers the same argument shows that formula (30) remains valid if all these points are excluded from the sum on the left-hand side.

Example.

$$\sum_{k=1}^{\infty}\frac{1}{k^2}=\frac{1}{2}\sum_{\substack{k=-\infty\\(k\neq 0)}}^{\infty}\frac{1}{k^2}=-\frac{\pi}{2}\operatorname*{Res}_{z=0}\frac{\cot(\pi z)}{z^2}=$$

$$=-\frac{\pi}{2}\operatorname*{Res}_{z=0}\left[\left(\frac{1}{\pi z}-\frac{\pi z}{3}+\ldots\right):z^2\right]=\frac{\pi^2}{6}$$

If instead of (29) we take the integral $\int_{(L_n)}\frac{f(z)}{\sin(\pi x)}\,dz$ it can readily be shown that, under the same assumptions, we obtain the formula

$$\sum_{k=-\infty}^{\infty}(-1)^n f(n)=-\pi\sum_{j}\operatorname*{Res}_{z=z_j}\frac{f(z)}{\sin(\pi z)}$$

5. Poisson Integral. The Poisson integral is used for deducing formulas which express the values of a harmonic function in a half-plane or circle in terms of its values taken on the boundary of the domain in question. To obtain the first formula let us consider the integral

$$\int_{-\infty}^{\infty}f(x)\left[\frac{1}{x-(\xi+i\eta)}-\frac{1}{x-(\xi-i\eta)}\right]dx\quad(\eta>0)\tag{31}$$

where $f(z)$ is an analytic function in the half-plane $\operatorname{Im} z\geqslant 0$ satisfying the condition

$$|f(z)|=o(|z|)\qquad(z\to\infty,\ \operatorname{Im} z\geqslant 0)$$

Then condition (16) is fulfilled for the integrand in integral (31) (why?). Computing the integral according to the scheme of Sec. 4 we obtain the value $2\pi i f(\xi+i\eta)$. Now performing the addition in the square brackets we arrive at the relation

$$f(\xi+i\eta)=\frac{\eta}{\pi}\int_{-\infty}^{\infty}\frac{f(x)}{(x-\xi)^2+\eta^2}\,dx$$

Separating the real part we obtain the formula

$$u(\xi,\ \eta)=\frac{\eta}{\pi}\int_{-\infty}^{\infty}\frac{u(x,\ 0)}{(x-\xi)^2+\eta^2}\,dx\qquad(\eta>0)$$

expressing the values of a harmonic function in the upper half-

plane in terms of its boundary values, the right member being termed a **Poisson integral.**

If in the square brackets in formula (31) we take the sum of the fractions instead of their difference, we similarly receive (check it up!) the formulas

$$u(\xi, \eta)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{(x-\xi)\,v(x, 0)}{(x-\xi)^2+\eta^2}dx,$$
$$v(\xi, \eta)=-\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{(x-\xi)\,u(x, 0)}{(x-\xi)^2+\eta^2}dx \qquad (\eta>0) \qquad (32)$$

each of which expresses the values of one of two conjugate harmonic functions in terms of the boundary values of the other. For formulas (32) to be valid it is sufficient that the condition

$$|f(z)|=o(1) \text{ (i.e. } f(z)\to 0) \text{ for } z\to\infty, \ \mathrm{Im}\, z\geqslant 0$$

hold. By the way, Jordan's lemma indicates that if $f(z)=f_1(z)\,e^{i\omega z}$ $(\omega>0)$, it is sufficient that $|f_1(z)|=o(|z|)$ (similarly, for formula (31) to hold in this case it is sufficient that $|f_1(z)|=o(|z|^2)$).

If we apply to the singular integral $\int_{-\infty}^{\infty}\frac{f(x)}{x-\xi}d\xi$ the technique indicated in Sec. 4 we obtain, as in the case of integral (32), the formulas

$$u(\xi, 0)=\frac{1}{\pi}\text{v.p.}\int_{-\infty}^{\infty}\frac{v(x, 0)}{x-\xi}dx, \quad v(\xi, 0)=-\frac{1}{\pi}\text{v.p.}\int_{-\infty}^{\infty}\frac{u(x, 0)}{x-\xi}d\xi \qquad (33)$$

These formulas can be rewritten in another form if we take into account that $\text{v.p.}\int\frac{dx}{x-\xi}=0$, which enables us to put down the relations

$$u(\xi, 0)=\frac{1}{\pi}\text{v.p.}\int_{-\infty}^{\infty}\frac{v(x, 0)-v(\xi, 0)}{x-\xi}dx$$
$$v(\xi, 0)=-\frac{1}{\pi}\text{v.p.}\int_{-\infty}^{\infty}\frac{u(x, 0)-u(\xi, 0)}{x-\xi}dx \qquad (34)$$

The point $x=\xi$ is no longer a singular point of the integrands in the last integrals, which are only singular because the limits of integration are infinite (why?).

Taking various concrete functions $f(z)$ we can find, by means of formulas (33) and (34), the values of many important integrals.

For instance, let us put $f(z)=e^{iz}$ and $\xi=0$. Then the first formula (33) yields the equality $\int_{-\infty}^{\infty}\frac{\sin x}{x}dx=\pi$.

Similar formulas can be established for integrals over a circle. To derive them we consider, instead of (31), the integral

$$\oint f(z)\left[\frac{1}{z-\zeta}-\frac{1}{z-a^2(\zeta^*)^{-1}}\right]dz \tag{35}$$

taken over the circle $z=ae^{i\varphi}\,(0\leqslant\varphi\leqslant 2\pi)$ where $\zeta=re^{i\theta}\,(r<a)$, the function $f(z)$ being analytic everywhere for $|z|\leqslant a$ and the points ζ and $a^2(\zeta^*)^{-1}$ being symmetric with respect to the circle $|z|=a$ (see Sec. 1.8). Integral (35) being equal to $2\pi i f(\zeta)$, we can make use of the simple transformation

$$(z-\zeta)(z-a^2(\zeta^*)^{-1})=(ae^{i\varphi}-re^{i\theta})\left(ae^{i\varphi}-\frac{a^2}{r}e^{i\theta}\right)=$$
$$=-\frac{a}{r}e^{i(\varphi+\theta)}[a^2-2ar\cos(\varphi-\theta)+r^2]$$

and thus obtain the formula

$$f(re^{i\theta})=\frac{1}{2\pi}\int_0^{2\pi}\frac{a^2-r^2}{a^2-2ar\cos(\varphi-\theta)+r^2}f(ae^{i\varphi})\,d\varphi\quad(0\leqslant r<a) \tag{36}$$

(let the reader check up the calculations). Here we can also separate the real part and receive a formula expressing the values of a harmonic function inside a circle in terms of its values on the circumference of the circle.

Putting $r=0$ in (36) we obtain the so-called **mean-value formula**

$$f(0)=\frac{1}{2\pi}\int_0^{2\pi}f(ae^{i\varphi})\,d\varphi \tag{37}$$

which shows that the value of an analytic (and therefore a harmonic) function at the centre of a circle is equal to its mean value on the circumference of the circle.

Taking in (35) the sum of the fractions instead of their difference and using formula (37) we derive the formula

$$f(re^{i\theta})=f(0)-\frac{iar}{\pi}\int_0^{2\pi}\frac{\sin(\varphi-\theta)}{a^2-2ar\cos(\varphi-\theta)+r^2}f(ae^{i\varphi})\,d\varphi$$

from which it is easy to deduce formulas similar to (32).

Let the reader consider the integral $\oint \frac{f(z)}{z-\zeta} d\zeta$ $(r=a)$ and apply formula (37) to prove that

$$f(ae^{i\theta}) = f(0) - \frac{1}{2\pi} \text{v.p.} \int_0^{2\pi} f(ae^{i\varphi}) \cot \frac{\varphi-\theta}{2} d\varphi$$

After that derive the formulas

$$\left.\begin{aligned} u(a, \theta) &= u(0) + \frac{1}{2\pi} \text{v.p.} \int_0^{2\pi} v(a, \varphi) \cot \frac{\varphi-\theta}{2} d\varphi, \\ v(a, \theta) &= v(0) - \frac{1}{2\pi} \text{v.p.} \int_0^{2\pi} u(a, \varphi) \cot \frac{\varphi-\theta}{2} d\varphi, \end{aligned}\right\} \quad (38)$$

(It should be noted that for $\theta=0$ the symbol $\text{v.p.} \int_0^{2\pi}$ must be understood as $\lim_{\varepsilon \to +0} \int_{\varepsilon}^{2\pi-\varepsilon}$.)

Let us consider an example. Putting $f(z) = z^n$ $(n=1, 2, 3, \ldots)$, which means that we take $u = \rho^n \cos n\theta$, $v = \rho^n \sin n\theta$, and choosing $a=1$ we obtain the useful formulas

$$\frac{1}{2\pi} \text{v.p.} \int_0^{2\pi} \sin n\varphi \cdot \cot \frac{\varphi-\theta}{2} d\varphi = \cos n\theta$$

and

$$\frac{1}{2\pi} \text{v.p.} \int_0^{2\pi} \cos n\varphi \cdot \cot \frac{\varphi-\theta}{2} d\varphi = -\sin n\theta$$

Taking advantage of the identity $\cot \frac{\varphi-\theta}{2} = -\frac{\sin\varphi + \sin\theta}{\cos\varphi - \cos\theta}$, passing to the integration from $-\pi$ to π and applying the rules of integration of even and odd functions we can rewrite the latter relations in the form

$$\text{v.p.} \int_0^{\pi} \frac{\cos n\varphi}{\cos\varphi - \cos\theta} d\varphi = \pi \frac{\sin n\theta}{\sin\theta}, \quad \text{v.p.} \int_0^{\pi} \frac{\sin n\varphi \cdot \sin\varphi}{\cos\varphi - \cos\theta} d\varphi = -\pi \cos n\theta$$

$$(n=1,2,3, \ldots)$$

It is readily proved that the first group of these formulas remains valid for $n=0$ as well. The formulas will be used in Sec. VII.5.7.

6. Behaviour of a Function at Infinity. Let $f(z)$ be a single-valued analytic function in the neighbourhood of $z=\infty$, that is for all sufficiently large $|z|$. Then the considerations of Sec. 2.4 imply that the function can be expanded in this neighbourhood into

Laurent's series (2.8):

$$f(z)=\sum_{k=-\infty}^{\infty} c_k z^k$$

(By the way, we can also write down a more general expansion (2.11).) The notion of a singularity at $z=\infty$ and the classification of singular points at infinity are introduced as follows. We make the substitution $z=\frac{1}{\zeta}$ and investigate the function thus obtained at the point $\zeta=0$. The properties of this function in the vicinity of the point $\zeta=0$ are then attributed, in the way indicated below, to the behaviour of the original function at $z=\infty$: we say that $f(z)$ has a pole at $z=\infty$ if the corresponding expansion (2.8) contains a finite number $n\geqslant 1$ of powers with positive exponents; similarly, we say that $f(z)$ has an essential singularity at $z=\infty$ if there is an infinitude of positive powers.

If the function $f(z)$ is bounded at infinity we can apply, after z has been replaced by $\frac{1}{\zeta}$, the argument given in the first paragraph of Sec. 2.5 and thus conclude that there are no powers with positive exponents in expansion (2.8). In particular, it follows that in this case there exists a finite value $f(\infty)$. Then we say that the function $f(z)$ is **analytic at $z=\infty$.**

There is a consequence of this consideration known as **Liouville's theorem:** *if a function $f(z)$ is single-valued, analytic and bounded throughout the z-plane it is identically equal to a constant.* Indeed, by the maximum modulus principle (Sec. 2.6), we have

$$\max_{|z|\leqslant R}|f(z)-f(\infty)|=\max_{|z|=R}|f(z)-f(\infty)|$$

for any $R>0$. If now $R\to\infty$, we see that the right-hand side, and therefore the left-hand side, tend to zero, which proves the theorem.

In its turn, Liouville's theorem implies that *if $f(z)$ is a single-valued analytic function in the whole z-plane except at its singular points (including possibly the point $z=\infty$) which are all poles, then $f(z)$ is a rational function* (let the reader prove the converse). In fact, these conditions indicate that there can only be a finite number of singular points because, if otherwise, it would follow from the Bolzano-Weierstrass theorem (see Sec. 2.6) that there is a non-isolated finite singular point or a non-isolated singularity at $z=\infty$, which contradicts the hypothesis. Now we can take Laurent's expansions of $f(z)$ at all its singular points and form the sum $g(z)$ of all principal parts of these series (it should be taken into account that in the expansion at $z=\infty$ the principal part is the one containing powers with positive exponents). Since these

singular points may only be poles, every principal part consists of a finite number of terms, and therefore their sum includes a finite number of algebraic summands, and thus the function $g(z)$ is rational. On the other hand, it can readily be shown (how?) that the difference $f(z)-g(z)$ is bounded throughout the z-plane and hence, by Liouville's theorem, it is equal to a constant, and hence the desired proposition has been proved.

In particular, we conclude that a *single-valued analytic function $f(z)$ without finite singular points which has a pole at $z=\infty$ is a polynomial.* It can be easily shown that irrational algebraic functions always have branch points. Therefore, among one-valued analytic functions defined in the whole complex plane and having a finite number of singular points (finite or infinite) the transcendental functions (and only they) possess essential singularities.

The residue of $f(z)$ at $z=\infty$ (at infinity) is the number $-c_{-1}$ taken from the expansion of type (2.8) at $z=\infty$. (It should be noted that the residue can be different from zero even when the function is analytic at $z=\infty$!)

We shall prove that *the sum of all residues of a single-valued analytic function $f(z)$ having a finite number of singular points in the entire z-plane is equal to zero.* Indeed, let a circle $|z|=R$ of a sufficiently large radius R (oriented in the positive direction) contain inside it all the finite singular points $z_1, z_2, \ldots, z_N$ of the function $f(z)$. Then, by the residue theorem, we have

$$\oint_{|z|=R} f(z)\,dz = 2\pi i \sum_{k=1}^{N} \operatorname*{Res}_{z=z_k} f(z)$$

On the other hand, taking expansion (2.8) in the neighbourhood of $z=\infty$ we see that the same integral is equal to $2\pi i c_{-1} = -2\pi i \operatorname*{Res}_{z=\infty} f(z)$. Equating the results and cancelling by $2\pi i$ we thus arrive at the above assertion which sometimes facilitates the computation of residues.

7. Logarithmic Residues. Let $f(z)$ be a one-valued analytic function in the vicinity of a (finite) point z_0 except possibly at this point itself at which there is either a zero or a pole. Then the function $\frac{f'(z)}{f(z)}$ has at $z=z_0$ an isolated singularity, and its residue

$$\operatorname*{Res}_{z=z_0} \frac{f'(z)}{f(z)} \tag{39}$$

at the point z_0 is called the **logarithmic residue** of the function $f(z)$ $\left(\text{this term is accounted for by the fact that } \frac{f'}{f} = (\operatorname{Ln} f)'\right)$.

If $f(z)$ has a zero (pole) of order n at the point z_0, its logarithmic residue (39) *is equal to n ($-n$).* Indeed, in both cases we can write

$$f(z) = (z-z_0)^{\nu} g(z)$$

where $\nu = n$ in the case of a zero and $\nu = -n$ for a pole, the function $g(z)$ is analytic at the point z_0 and $g(z_0) \neq 0$ (why?). It follows that

$$\frac{f'(z)}{f(z)} = \frac{\nu g(z) + (z-z_0) g'(z)}{(z-z_0) g(z)}$$

and thus, computing the residue of this function, we obtain, according to formula (5), the relation

$$\operatorname*{Res}_{z=z_0} \frac{f'(z)}{f(z)} = \frac{\nu g(z_0) + (z_0-z_0) g'(z_0)}{g(z_0) + (z_0-z_0) g'(z_0)} = \nu$$

which is what we set out to prove.

Cauchy's residue theorem (9) applied to the function $\frac{f'}{f}$ and the assertion proved above imply the following

Theorem. *Let $f(z)$ be a single-valued analytic function which is different from zero everywhere on a closed positively oriented contour (L) and inside it except possibly at a finite number of zeros and poles lying in the interior of (L). Then the total number N of zeros and the total number P of poles (each zero and each pole is counted a number of times equal to its order) are connected by the formula*

$$\oint_{(L)} \frac{f'(z)}{f(z)}\, dz = 2\pi i\,(N-P) \tag{40}$$

The left-hand side of formula (40) is equal to

$$\oint_{(L)} d \operatorname{Ln} f(z) = \Delta_{(L)} \operatorname{Ln} f(z) = \Delta_{(L)} [\ln |f(z)| + i \operatorname{Arg} f(z)] = i \Delta_{(L)} \operatorname{Arg} f(z)$$

where the symbol $\Delta_{(L)}$ designates the increment gained when the point z traverses the contour (L); therefore it follows that

$$N - P = \frac{1}{2\pi} \Delta_{(L)} \operatorname{Arg} f(z) \tag{41}$$

The right-hand side of the last relation is equal to the number of circuits made by the point $w = f(z)$ (in the positive direction) about the origin in the w-plane when the point z describes the contour (L). Formula (41) is referred to as the **argument principle.**

It should be noted that the condition that the function $f(z)$ is analytic on the contour (L) itself (i.e. in a curvilinear strip containing this contour) which has been used for proving the

argument principle can be in fact replaced by a less restrictive requirement that the function does not vanish on the contour and is continuous on it. To prove the argument principle under this weaker hypothesis we can apply formula (41) to a contour (L') approximating (L) from inside (see Fig. 45) and then pass to the limit for $(L') \to (L)$.

8. Rouché's Theorem. *Let $f(z)$ and $g(z)$ be single-valued analytic functions everywhere inside a closed contour (L) and on it, and let the strict inequality $|g(z)| < |f(z)|$ be fulfilled on the contour.*

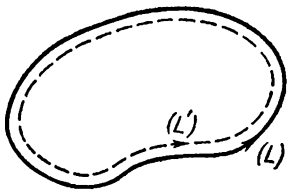


Fig. 45

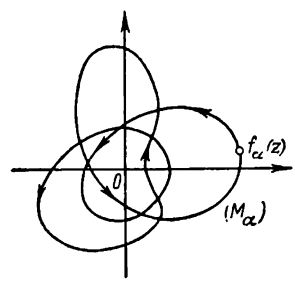


Fig. 46

Then the functions $f(z)$ and $f(z)+g(z)$ have the same number of zeros inside (L) (here and henceforward each zero is counted according to its multiplicity).

To prove this proposition (known as **Rouché's theorem**) we apply the argument principle to the function $f_\alpha(z) = f(z) + \alpha g(z)$ where α is a real parameter, $0 \leqslant \alpha \leqslant 1$. If the point z describes (L) in the positive direction, the point $w = f_\alpha(z)$ traverses an oriented closed contour (M_α) and makes N_α circuits about the origin in the w-plane (see Fig. 46 corresponding to $N_\alpha = 3$). By formula (41) (in which for this case we have $P = 0$), the number of circuits N_α is equal to the number of zeros of the function $f_\alpha(z)$ inside (L). If the parameter α receives a small increment the contour (M_α) undergoes a small deformation and therefore the number of circuits N_α remains unchanged. The number N_α may only change if the contour (M_α) passes through the origin for a certain value of α, i.e. if there is a point $\tilde{z} \in (L)$ such that

$$f_\alpha(\tilde{z}) = f(\tilde{z}) + \alpha g(\tilde{z}) = 0$$

But then $|f(\tilde{z})| - \alpha |g(\tilde{z})| \leqslant |g(\tilde{z})|$, which contradicts the hypothesis and hence this case is impossible. Consequently, N_α retains constant value when α ranges within the interval $0 \leqslant \alpha \leqslant 1$. In particular, it follows that $N_0 = N_1$ which is what we set out to prove.

We suggest that the reader carefully study the above proof because it involves two important methods widely used in modern

applied mathematics. The first method is based on introducing *an auxiliary parameter* (note that the original problem did not contain any parameter), which is performed in an artificial manner and is aimed at creating a continuous passage from one object to another. The other method is *continuous extension with respect to the parameter*: to prove that a property we are interested in (in the above proof this property is expressed by the equality $N_\alpha = N_0$) is valid for a finite interval of variation of the parameter we show that it takes place for a concrete point of the interval and then prove that in a small neighbourhood of every point of the interval it either holds everywhere or does not hold. An argument of this type enables us to pass from an assertion of local character (with respect to a parameter) to the corresponding assertion of global character.

Among the consequences of Rouché's theorem we mention here **the fundamental theorem of algebra** (e.g. see [107], Sec. VIII.8). We shall formulate it as follows: *every polynomial*

$$a_0 z^n + a_1 z^{n-1} + a_2 z^{n-2} + \ldots + a_n \quad (a_0 \neq 0)$$

of degree n has exactly n complex roots (counting multiplicities). To prove the theorem we put $f(z) = a_0 z^n$, $g(z) = a_1 z^{n-1} + a_2 z^{n-2} + \ldots + a_n$, take a circle $|z| = R$ of a sufficiently large radius R as contour (L) and then apply Rouché's theorem (let the reader complete the proof).

9. Zeros Dependent on a Parameter. Let $f_\alpha(z) = f(z, \alpha)$ be an entire analytic function of a complex variable z, i.e. a single-valued analytic function in the whole z-plane, which also depends on a parameter α. (We can similarly consider a function with an arbitrary domain of analyticity which may also depend on a parameter.) Suppose that in every finite region of the plane this function depends on α continuously in the sense of uniform deviation with respect to z, which means that for any chosen α and a sufficiently small $\Delta\alpha$ the uniform deviation of $f_{\alpha+\Delta\alpha}(z)$ from $f_\alpha(z)$ (i.e. the quantity $\max_z |f_{\alpha+\Delta\alpha}(z) - f_\alpha(z)|$) is arbitrarily small. Then we can assert that *the zeros of the function* $f_\alpha(z)$, *that is the roots of the equation*

$$f_\alpha(z) = 0 \tag{42}$$

are also continuous functions of α. The proof of this assertion and the detailed discussion are given below.

Let us first suppose that $\tilde{z}$ is a simple (i.e. not multiple) root of equation (42) for a value $\tilde{\alpha}$ of the parameter, and $f_{\tilde{\alpha}}(z) \not\equiv 0$. Denote by $(\tilde{L})$ an arbitrary sufficiently small circle with centre at $\tilde{z}$. The assertion in the second paragraph of Sec. 2.6 implies that $f_{\tilde{\alpha}}(z) \neq 0$ on $(\tilde{L})$, and therefore we also have $\min_{z \in (\tilde{L})} |f_{\tilde{\alpha}}(z)| =$

$=h>0$. If $\Delta\alpha$ is sufficiently small we can assert, by the hypothesis, that, at the points of the circle $(\tilde{L})$, the modulus of the difference $f_{\tilde{\alpha}+\Delta\alpha}(z)-f_{\tilde{\alpha}}(z)$ is less than a fixed number h. But then we can take this difference as function $g(z)$ and apply Rouché's theorem, which implies that the function $f_{\tilde{\alpha}+\Delta\alpha}(z)$ has exactly one simple root inside $(\tilde{L})$ because this is true for the function $f_{\tilde{\alpha}}(z)$. Thus, a sufficiently small variation of α results in an arbitrarily small variation of the root $\tilde{z}$ of equation (42).

Let us now consider a root $\tilde{z}$ of multiplicity k. Applying the same argument we conclude that sufficiently small variations of α result in the appearance of a number of roots of equation (42) lying arbitrarily close to the former root $\tilde{z}$, the sum of the multiplicities of these new roots being equal to k. Thus the equation may retain one root of multiplicity k or have k simple roots or, finally, there may occur some intermediate cases. This again indicates that a multiple root of order k should be regarded as k coincident roots which may separate when the parameter is changed.

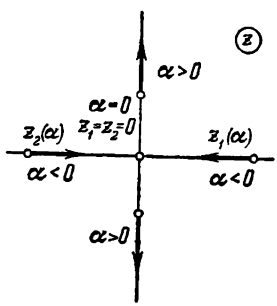


Fig. 47

It should be noted that the assertion proved above is by no means trivial. If we only use real numbers it does not necessarily hold in the general case. For instance, if the real parameter α entering into the equation $x^2+\alpha=0$ increases, passes through zero and becomes positive we see that, from the point of view of real analysis, its two roots $\pm\sqrt{-\alpha}$ are distinct for $\alpha<0$, coincide for $\alpha=0$ and disappear for $\alpha>0$. Thus, for the above assertion to be true we must resort to the set of complex numbers. This assertion can be interpreted as *stability principle for the roots of an equation with respect to the variation of the equation*.

If α is a real parameter then, from what has been proved, we see that the roots of equation (42) describe continuous lines in the z-plane when the parameter α is continuously changed. (Here we do not consider a more complicated case when for certain values of α the identity $f_\alpha(z)\equiv 0$ holds.) Let us discuss some problems related to continuation of these lines. For instance, suppose that, for a value $\alpha=\alpha_0$, two simple roots $z_1(\alpha)$ and $z_2(\alpha)$ merge into a twofold root and then separate. Then, as a rule, it is impossible to decide which of the roots moves along one path and which along another, and thus, after the parameter has passed the value α_0, we should supply them with new numbers which are irrelevant to the former indices. For example, see

Fig. 47 which represents the behaviour of the roots of the equation $x^2+\alpha=0$ mentioned above. When the parameter assumes the value $\alpha=0$ the roots merge, and we cannot say which of them goes upward and which goes downward after α has become positive.

When the roots of equation (42) change as the parameter is continuously varied, one or several roots may recede to infinity for a finite value $\alpha=\alpha_0$ of the parameter. Then we can say that they disappear because we only consider finite roots. After α has passed the value α_0 these roots "return from infinity". Hence, due to the above facts, the total number of roots in the plane may decrease for some separate values of the parameter but this does not contradict the stability of the roots established above.

Our discussion enables us to formulate the general rule for determining the number of roots of equation (42) in a given domain (G) of the z-plane. Let us first suppose that this domain is finite and that for a given value α_0 the number N of roots in (G) is known. Then, if, for every α taken from a given interval, there are no roots on the boundary (Γ) of the domain (G) we can assert that for all values of α belonging to this interval the number of roots in (G) is equal to N (why?). But if for a value $\tilde{\alpha}$ of the parameter there appear one or more roots on (Γ) we must find out whether these roots move inward or outward when α is changed and thus determine the variation of the number of roots inside (G) as the parameter α passes through $\tilde{\alpha}$. If the domain (G) is infinite we must also take into account that some roots located in (G) may tend to infinity or return from infinity; to examine this case it is advisable to resort to the asymptotic expansion of the function f for $z\to\infty$.

There is an important particular case when we have an analytic function $f(z)$ which is real, i.e. assumes real values for real z. Then its Taylor expansion in powers of $z-z_0$ must have real coefficients for real z_0, and therefore for the conjugate complex values of z the function also takes on conjugate complex values, that is $f(z^*)\equiv[f(z)]^*$ (to establish this fact one must use the properties of conjugate complex numbers; e.g. see [107], Sec. VIII.3). It follows that if $f(c)=0$, we have $f(c^*)=0$, and if $f'(c)=0$ then we also have $f'(c^*)=0$ (because the derivative of a real function is also real), etc. Hence, if a real analytic function possesses an imaginary zero it must also have the conjugate complex zero of the same multiplicity. (It is well known that, in particular, the polynomials with real coefficients possess this property; e.g. see [107], Sec. VIII.8.)

It follows that if the left-hand side of equation (42) is a real analytic function, its simple real roots cannot leave the real axis

without merging into multiple roots when the parameter changes (see Fig. 47).

The lines along which the roots $z_k(\alpha)$ move can be constructed numerically by applying the method of continuous extension, described in the second half of Sec. IV.4.11, after we pass from complex equation (42) to the corresponding system of two real equations with $\mathrm{Re}\, z$ and $\mathrm{Im}\, z$ as unknowns. When roots of the equation merge the corresponding system of differential equations degenerates, and therefore, after they have merged, it is advisable to use the expansions of the roots in powers of the increment of the parameter (see Sec. 2.8) and then continue these expansions by solving the differential equations. Considering the direction field of the corresponding differential equation we can also investigate the way in which the trajectories of the roots intersect the contour (Γ) (this phenomenon was discussed above).

10. Zeros of Polynomials. We shall apply the considerations of Sec. 9 to studying the dependence of the zeros of a polynomial

$$P_\alpha(z) = a_0(\alpha) z^n + a_1(\alpha) z^{n-1} + \ldots + a_n(\alpha) \qquad (43)$$

on the real parameter α, on condition that the coefficients of the polynomial are continuous functions of the parameter. As was shown at the end of Sec. 8, if $a_0(\alpha) \neq 0$ there are exactly n zeros (roots) $z_1(\alpha)$, $z_2(\alpha)$, ..., $z_n(\alpha)$ among which some may coincide. When α is varied these zeros describe n trajectories in the z-plane (here we must take into account the stipulations concerning the arbitrariness in numbering these trajectories after they have intersected).

If the degree of polynomial (43) is reduced by $k \geqslant 1$ for a value $\alpha = \tilde{\alpha}$ of the parameter, that is if $a_0(\tilde{\alpha}) = a_1(\tilde{\alpha}) = \ldots = a_{k-1}(\tilde{\alpha}) = 0$, $a_k(\tilde{\alpha}) \neq 0$, then, for this value of α, only $n-k$ zeros remain in the z-plane. Hence, for this value of the parameter α, exactly k trajectories are at infinity. If, after α has passed the value $\tilde{\alpha}$, the coefficient $a_0(\alpha)$ is again nonzero, the trajectories "return from infinity".

Now suppose that $a_0(\alpha) \neq 0$ for all the values of α under consideration, and we are interested in the number $N(\alpha)$ of the zeros of polynomial (43) lying in the "right half-plane", i.e. in the domain $\mathrm{Re}\, z > 0$, the value N_0 of $N(\alpha)$ corresponding to a given value $\alpha = \alpha_0$ of the parameter being known. (Problems of this kind appear in the stability theory; e.g. see [107], Sec. XV.22.) We shall suppose, for simplicity, that the coefficients of polynomial (43) are real although the case of complex coefficients can be treated in like manner.

Since the zeros are continuous functions of α and cannot approach infinity (or "return from infinity") because $a_0(\alpha) \neq 0$, the number

$N(\alpha)$ may only change for a value $\alpha=\tilde{\alpha}$ if at least one root $\tilde{z}$ falls on the boundary of the domain we are interested in, that is if $\operatorname{Re}\tilde{z}=0$. Putting $\tilde{z}=i\tilde{y}$ (where $\tilde{y}$ is real), substituting this expression into (43) and separating the real and imaginary parts we arrive at the two equalities

$$a_0(\tilde{\alpha})\tilde{y}^n - a_2(\tilde{\alpha})\tilde{y}^{n-2} + a_4(\tilde{\alpha})\tilde{y}^{n-4} - \ldots = 0$$

and

$$a_1(\tilde{\alpha})\tilde{y}^{n-1} - a_3(\tilde{\alpha})\tilde{y}^{n-3} + a_5(\tilde{\alpha})\tilde{y}^{n-5} - \ldots = 0 \qquad (44)$$

which must hold simultaneously. Eliminating $\tilde{y}$ from (44) we obtain a relationship between the coefficients specifying the values of α for which $N(\alpha)$ may change. If some values $\tilde{\alpha}$ and $\tilde{y}$ satisfying both equalities (44) are found then, when α increases and passes through this value $\tilde{\alpha}$, the number $N(\alpha)$ increases or decreases by unity (because the zero passes through the point $i\tilde{y}$) depending on whether the quantity

$$\operatorname{Re}\frac{dz}{d\alpha} = -\operatorname{Re}\frac{\partial P_\alpha/\partial\alpha}{\partial P_\alpha/\partial z}\Big|_{\alpha=\tilde{\alpha},\ z=i\tilde{y}} \qquad (45)$$

is positive or negative.

If a polynomial involves several real parameters, for instance, $\alpha_1, \alpha_2, \ldots, \alpha_k$, we can similarly consider the domains of the k-dimensional space of the parameters which correspond to different values of $N(\alpha_1, \alpha_2, \ldots, \alpha_k)$. These domains can be visualized for $k=1$ and $k=2$. They are separated by $(k-1)$-dimensional surfaces whose equations are obtained from the conditions that the polynomial has pure imaginary zeros on them. (It should be noted that the number 0 is regarded as being pure imaginary.) We can similarly determine these domains for an arbitrary entire function dependent on parameters, but in this case there may appear additional surfaces in the space of the parameters from which zeros with $\operatorname{Re} z > 0$ move away to approach infinity. In recent years a number of articles have been published where the approach described above is applied to various classes of polynomials and the so-called **quasi-polynomials** which are functions of the form

$$\sum_{k=1}^{m} a_k z^{n_k} e^{\lambda_k z} \qquad (46)$$

(the numbers n_k are nonnegative integers).

The case when all the zeros of a polynomial have negative real parts, i.e. are in the left half-plane, is of particular importance for the stability theory. A polynomial of that type is called **stable** and is also referred to as a **Hurwitz polynomial** after A. Hurwitz

(1859-1919), a German mathematician, who proved in 1895 the following theorem (for the proof we refer the reader to special monographs; e.g. see [19]):

Hurwitz theorem. *For a polynomial*

$$P(z)=a_0z^n+a_1z^{n-1}+\ldots+a_{n-1}z+a_n \qquad (a_0>0) \qquad (47)$$

with real coefficients to be stable it is necessary and sufficient that for all $k=1, 2, \ldots, n$ *the condition*

$$\Delta_k=\begin{vmatrix} a_1 & a_3 & a_5 & \ldots & a_{2k-1} \\ a_0 & a_2 & a_4 & \ldots & a_{2k-2} \\ a_{-1} & a_1 & a_3 & \ldots & a_{2k-3} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{2-k} & a_{4-k} & a_{6-k} & \ldots & a_{2k-k} \end{vmatrix}>0$$

hold where the coefficients a_j *with* $j<0$ *or* $j>n$ *are put equal to zero.* A determinant of the type of Δ_k is referred to as a **Hurwitz determinant.** This theorem expresses the **Hurwitz criterion** for stability of a polynomial which is particularly convenient when the degree of the polynomial is not high.

We also draw attention to the following necessary condition: *if polynomial* (47) *with real coefficients is stable all its coefficients are positive.* This immediately follows from the possibility of factoring a real polynomial into real linear and quadratic factors (e.g. see [107], Sec. VIII.8). Indeed, if the polynomial is stable, the coefficients of these factors are positive (why?), and therefore, after the brackets have been opened, we arrive at the desired condition. It should be noted that this condition is sufficient only for $n=1$ and $n=2$ (check it up!) and is not sufficient in the general case. But it can be proved that if this condition is fulfilled and, besides, $\Delta_{n-1}>0$, $\Delta_{n-3}>0$, etc., polynomial (47) is stable. The latter assertion facilitates the application of the Hurwitz criterion.

The general problem of determining the number of zeros with positive real parts for a given entire function is known as the **Routh-Hurwitz problem.** A. Hurwitz proved that if, for polynomial (47), the determinants $\Delta_1, \Delta_2, \ldots, \Delta_n$ are different from zero, the number of zeros with positive real parts is equal to the number of changes of the sign in the numerical sequence $1, \Delta_1, \frac{\Delta_2}{\Delta_1}, \frac{\Delta_3}{\Delta_2}, \ldots, \frac{\Delta_n}{\Delta_{n-1}}$.

Consider an example. Let the parameter α entering into the polynomial $z^3+3z^2+3z+\alpha$ be real. Applying equations (44) to this polynomial we obtain (check it up!) the values $\tilde{\alpha}=0$ and $\tilde{\alpha}=9$ which break up the α-axis into three parts. Determining the sign of expression (45) (let the reader perform the calculations)

we conclude that the number N of zeros with positive real parts decreases by unity when the point α moves in the positive direction of the α-axis and passes through the point $\alpha = 0$, and increases by two when it passes through $\alpha = 9$. There are no such zeros for $\alpha = 1$ (why?), and therefore we see that $N = 1$ for $-\infty < \alpha < 0$, $N = 0$ for $0 < \alpha < 9$ (this is the interval of stability for the polynomial in question) and $N = 2$ for $9 < \alpha < \infty$. (Let the reader obtain the same result by applying the Hurwitz theorem.)

Another method for solving the Routh-Hurwitz problem is based on the direct application of the argument principle (Sec. 7). We shall demonstrate this method by considering a polynomial

$$P(z) = a_0 z^n + a_1 z^{n-1} + \ldots + a_n$$

although it can also be extended to many other classes of functions. Suppose that the polynomial $P(z)$ has no zeros on the imaginary axis. Let N be the sought-for number of zeros of the polynomial in the right half-plane. Then, by the argument principle, for a sufficiently large R the image under the mapping $w = P(z)$ of the closed contour (L_R) (shown in Fig. 48*a*) is a closed contour (L'_R) (see Fig. 48*b*) which makes N circuits about the origin in the w-plane, i.e. $\frac{1}{2\pi} \Delta_{(L)} \operatorname{Arg} P(z) = N$.

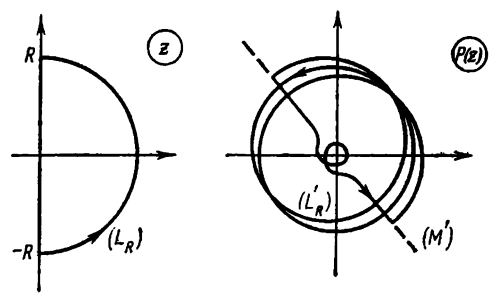


Fig. 48

But, for large $|z|$ we have the asymptotic expression $P(z) \sim a_0 z^n$, and therefore the increment of $\operatorname{Arg} P(z)$ corresponding to the semicircle of sufficiently large radius R is $\Delta \operatorname{Arg} P(z) \approx n\pi$. Consequently, if we denote the increment of $\operatorname{Arg} P(z)$, when the point z travels from $-\infty$ to $+\infty$ along the y-axis, by

$$\mu = \frac{1}{2\pi} \Delta_{-\infty < y < \infty} \operatorname{Arg} P(iy) \tag{48}$$

we can write

$$N = \frac{n}{2} - \mu \tag{49}$$

(Let the reader explain the appearance of the minus sign in (49) and also verify that Fig. 48*b* corresponds to the case $N = 4$, $n = 5$, $\mu = -\frac{3}{2}$.) Formula (49) gives the solution of the problem. The value of μ determined by formula (48) can be computed numeri-

cally on an electronic digital computer (for large $|y|$ it is convenient to introduce the new variable $\eta = \frac{1}{y}$ and pass to the case $\eta \to 0$) or by an approximate point-by-point plotting of the curve (M') (see Fig. 48*b*) after the necessary values of $P(iy)$ have been computed.

In particular, putting $N = 0$ we obtain the relation $\mu = n/2$ which is a necessary and sufficient condition for stability of a polynomial referred to as the **Mikhailov criterion.**

If $P(z)$ is a stable polynomial and y runs over the real values from $-\infty$ to ∞, the quantity $\operatorname{Arg} P(iy)$ varies monotonically (why?) and increases by πn. Consequently, the path of the point $P(iy)$ intersects the real axis, and also the imaginary axis, n times (one of the intersections can be attained in the limit when $y \to \pm\infty$). But this means that if we write $P(iy)$ in the form

$$P(iy) = Q(y) + iR(y) \tag{50}$$

where Q and R are real polynomials, the zeros of these polynomials are real and simple. Besides, there is exactly one zero of R between any two neighbouring zeros of Q and vice versa. Conversely, if the zeros of the polynomials Q and R entering into formula (50) possess the above properties (which can be easily tested by means of an electronic computer), all the zeros of the polynomial $P(z)$ lie either in the left half-plane or in the right half-plane. To find out which of these possibilities is realized we should determine (this can easily be done by numerical methods) in what direction the point $P(iy)$ turns about the origin. (Let the reader prove the converse on the basis of the hypotheses concerning the zeros by investigating the behaviour of the point $P(iy)$ when y increases from $-\infty$ to ∞.)

11. Resultant of Two Polynomial Equations. Here we shall discuss a problem related to the subject matter of Sec. 10 which is encountered in various divisions of mathematics. Suppose we have two polynomials $P(z)$ and $Q(z)$. It is required to eliminate z from the system of two equations

$$P(z) = 0, \quad Q(z) = 0 \tag{51}$$

or, which is the same, to derive a necessary and sufficient condition on the coefficients of the polynomials for equations (51) to have at least one root in common (why are these two formulations equivalent?).

For simplicity, we shall take a polynomial $P(z)$ of the third degree and a polynomial $Q(z)$ of the second degree. Then equations (51) are of the form

$$a_0z^3 + a_1z^2 + a_2z + a_3 = 0, \quad b_0z^2 + b_1z + b_2 = 0 \tag{52}$$

(By the way, the result obtained for this simple system shows what answer must be given in the general case of polynomials of arbitrary degrees.) Suppose that equations (52) have a common solution $z=z_0$. Substitute z_0 into the first and second equations and then multiply the first expression obtained by z_0 and the second by z_0^2 and z_0. Thus we obtain the equalities

$$\begin{aligned}
a_0z_0^4+a_1z_0^3+a_2z_0^2+a_3z_0 &= 0\\
a_0z_0^3+a_1z_0^2+a_2z_0+a_3\cdot 1 &= 0\\
b_0z_0^4+b_1z_0^3+b_2z_0^2 &= 0\\
b_0z_0^3+b_1z_0^2+b_2z_0 &= 0\\
b_0z_0^2+b_1z_0+b_2\cdot 1 &= 0
\end{aligned}$$

which can be regarded as a system of five homogeneous algebraic equations of the first degree with respect to the five unknown quantities z_0^4, z_0^3, ..., 1. These quantities are not all equal to zero, and therefore the determinant of the system must vanish (e.g. see [107], Sec. VI.6):

$$\begin{vmatrix}
a_0 & a_1 & a_2 & a_3 & 0\\
0 & a_0 & a_1 & a_2 & a_3\\
b_0 & b_1 & b_2 & 0 & 0\\
0 & b_0 & b_1 & b_2 & 0\\
0 & 0 & b_0 & b_1 & b_2
\end{vmatrix}=0 \qquad (53)$$

This determinant, which can be similarly formed for polynomials $P(z)$ and $Q(z)$ of arbitrary degrees (its degree is equal to the sum of the degrees of the polynomials P and Q), is called **the resultant of polynomial equations** (51). We have thus shown that equality (53) is necessary for equations (52) to have a finite root in common. If there is a common root at infinity, that is if $a_0=b_0=0$ (see Sec. 10), it is clear that relation (53) is also fulfilled. It can be proved that *the condition that the resultant is equal to zero is not only necessary but also sufficient for two polynomial equations to have a common (finite or infinite) zero* (e.g. see [76]).

The resultant of equations $P(z)=0$ and $P'(z)=0$ where $P(z)$ is a polynomial of degree n is called the **discriminant** *) of the equation $P(z)=0$. *For $P(z)$ to have a multiple root (or a root at infinity) it is necessary and sufficient that the discriminant be equal to zero* (why?). (Let the reader form the discriminant of the quadratic equation $az^2+bz+c=0$ and compare it with the expres-

*) The discriminant of a polynomial equation $P(z)=0$ of degree n is sometimes defined as being $(-1)^{\frac{n(n-1)}{2}}$ times the resultant of the equation and its derived equation $P'(z)=0$.— *Tr.*

sion b^2-4ac which, as is known, is also termed a discriminant in elementary mathematics; see also footnote on page 131.)

Using the resultant we can (at least in principle) pass from a system of equations

$$P(z_1, z_2)=0, \quad Q(z_1, z_2)=0$$

where P and Q are polynomials to the corresponding algebraic equation $D(z_1)=0$ in one unknown, but in concrete problems this may lead to very complicated calculations. The resultant can be similarly applied to constructing solutions of a system of equations

$$P(z_1, z_2, z_3)=0, \quad Q(z_1, z_2, z_3)=0$$

in the vicinity of the point $z_1=z_2=z_3=0$ by using the method of Newton's polygon (see Sec. 2.8) for the obtained equation $D(z_1, z_2)=0$.

12. Meromorphic Functions. A single-valued analytic function defined in the whole z-plane and having only poles as its finite singular points is called a **meromorphic** function. A function of this kind can have only a finite number of poles in every finite part of the plane (why?), and therefore if there is an infinitude of poles they must accumulate at infinity so that the point at infinity is a **non-isolated singular point** of the function. The function $\tan z$ whose poles are $z=\frac{\pi}{2}+k\pi$ $(k=0, \pm 1, \pm 2, \ldots)$ is a typical example of a meromorphic function with non-isolated singularity at $z=\infty$.

If a meromorphic function $f(z)$ has only a finite number of poles we can take the sum $s(z)$ of the principal parts of the corresponding Laurent expansions at all the poles and thus represent the function in the form $f(z)=s(z)+[f(z)-s(z)]$ as a sum of a rational function and an entire function. It turns out that a similar representation is possible for a wide class of meromorphic functions with infinite number of poles but in this case $s(z)$ is a sum of an infinite series of rational functions. Here we shall only formulate the simplest theorem related to these questions.

Suppose that for a meromorphic function $f(z)$ with an infinite number of poles it is possible to choose a sequence of closed contours (L_1), (L_2), ..., (L_n), ..., each of which is contained inside the subsequent contour, such that the following conditions are fulfilled:

1. $\min\limits_{z\in(L_n)} |z| \xrightarrow[n\to\infty]{} \infty$, i.e. the contours recede to infinity as $n \longrightarrow \infty$;
2. $L_n = O\left(\min\limits_{z\in(L_n)} |z|\right)$, which means that the length of the contour (L_n) is of order not higher than its minimum distance from the origin;

3. $\sup_n \max_{z \in (L_n)} |f(z)| < \infty$ ("sup" means the least upper bound; e.g. see [107], Sec. IV.19), that is the function $f(z)$ is uniformly bounded on all the contours (L_n). *Then the function $f(z)$ can be represented in the form*

$$f(z) = \sum_{n=1}^{\infty} [s_n(z) - s_n(z_0)] + f(z_0), \tag{54}$$

where $s_n(z)$ is the sum of the principal parts of the Laurent expansions of the function $f(z)$ at all its poles lying between (L_{n-1}) and (L_n) for $n > 1$ and inside (L_1) for $n = 1$ and z_0 is an arbitrary non-singular point of the function $f(z)$. Series (54) is uniformly convergent in every finite part of the z-plane.

In order to prove this theorem we fix two non-singular points z_0 and z and consider the auxiliary function

$$F(\zeta) = \frac{f(\zeta)}{\zeta - z} - \frac{f(\zeta)}{\zeta - z_0} = \frac{(z - z_0) f(\zeta)}{(\zeta - z)(\zeta - z_0)} \tag{55}$$

of the variable ζ. This function has a pole at the point $\zeta = z$ with residue $f(z)$ and a pole at $\zeta = z_0$ with residue $-f(z_0)$ (why?) and, besides, the same poles $\zeta = z_1, z_2, z_3, \ldots$, as the function $f(\zeta)$. Applying formula (6) to calculating the residue of $F(\zeta)$ at the point $\zeta = z_k$ and taking into account that the expansion of $f(\zeta)$ contains terms of the form $\frac{A_i}{(\zeta - z_k)^p}$ we find that

$$\operatorname*{Res}_{\zeta = z_k} F(\zeta) = -f_k(z) + f_k(z_0)$$

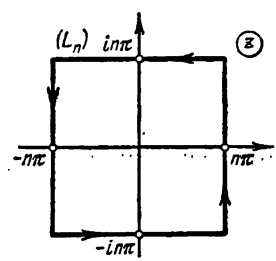


Fig. 49

where $f_k(z)$ is the principal part of the expansion of the function $f(z)$ at the point $z = z_k$ (check up the calculations!). Therefore, by Cauchy's residue theorem, we obtain

$$\oint_{(L_N)} F(\zeta)\, d\zeta = 2\pi i \left\{ f(z) - f(z_0) - \sum_{n=1}^{N} [s_n(z) - s_n(z_0)] \right\}$$

Now, if $N \to \infty$, formula (55), inequality (2.3) and the conditions of the theorem imply that the left-hand member tends to zero, the passage to the limit being uniform with respect to z in every finite part of the z-plane. This proves the assertion of the theorem.

Let us apply this theorem to the function $f(z) = \tan z$ putting $z_0 = 0$ and taking as $(L_1), (L_2), \ldots$ the squares of the form shown in Fig. 49. We have

$$\tan(\pm n\pi + iy) = \tan iy = i \tanh y$$

and

$$|\tan(x \pm in\pi)| = \left|\frac{\tan x \pm i \tanh n\pi}{1 \mp i \tanh n\pi \tan x}\right| \leqslant \frac{(|\tan x|+1)^2}{1+\tanh^2 n\pi \tan^2 x} \leqslant$$
$$\leqslant \frac{(|\tan x|+1)^2}{1+\tanh^2 \pi \tan^2 x} \leqslant \text{const}$$

(check it up!), and therefore Condition 3 of the theorem is fulfilled (the validity of the other conditions is obvious). But the function $\tan z$ has the poles $z=\frac{\pi}{2}+k\pi$ for which the principal parts of the Laurent expansions are $-\left[z-\left(\frac{\pi}{2}+k\pi\right)\right]^{-1}$ (check it up!), and consequently formula (54) yields

$$\tan z = \sum_{k=-\infty}^{\infty}\left[\frac{-1}{z-\left(\frac{\pi}{2}+k\pi\right)} - \frac{1}{\frac{\pi}{2}+k\pi}\right] \tag{56}$$

Combining the terms with $k=0$ and $k=-1$, the terms with $k=1$ and $k=-2$, etc. (strictly speaking, formula (54) indicates that the summation should be carried out exactly in this way) we receive

$$\tan z = -2z\sum_{k=0}^{\infty}\frac{1}{z^2-\left(\frac{\pi}{2}+k\pi\right)^2}$$

If in formula (56) we first perform the summation from $k=-n$ to $k=n$ and then pass to the limit for $n\to\infty$, it is readily seen that the sum of the second terms tends to zero, and thus

$$\tan z = \lim_{n\to\infty}\sum_{k=-n}^{n}\frac{-1}{z-\left(\frac{\pi}{2}+k\pi\right)} \tag{57}$$

The series $\sum_{k=-\infty}^{\infty}\frac{-1}{z-\left(\frac{\pi}{2}+k\pi\right)}$ is divergent, and its sum determined by formula (57) is understood in the sense of the *principal value* (e.g. cf. [107], Sec. XIV.19).

Replacing in formula (57) z by ζ, integrating with respect to ζ from 0 to z and then taking antilogarithms we obtain another interesting formula:

$$\cos z = \lim_{n\to\infty}\prod_{k=-n}^{n}\left(1-\frac{z}{\frac{\pi}{2}+k\pi}\right)$$

(the symbol $\prod_{k=m}^{n}$ designates a product of terms with index k ranging from m to n; **infinite products** involving the symbols $\prod_{k=1}^{\infty}$ or

$\prod_{k=-\infty}^{\infty}$ are defined by analogy with series, and the principal value of a divergent infinite product is also defined in a similar way; it should also be noted that the product $\prod_{k=-\infty}^{\infty}\left(1-\frac{z}{\frac{\pi}{2}+k\pi}\right)$ is divergent and the above formula determines its principal value).

Some other entire functions $f(z)$ can be represented as infinite products by performing similar operations on expansions of the meromorphic function $\frac{f'(z)}{f(z)}$ into series. For instance, it can be shown (but we do not present the proof here) that the meromorphic function $\frac{\Gamma'(z)}{\Gamma(z)}$ satisfies the conditions of the above theorem on expansion (54), and therefore, putting $z_0=1$, we obtain (check it up!) the relation

$$\frac{\Gamma'(z)}{\Gamma(z)}=\sum_{n=0}^{\infty}\left(-\frac{1}{z+n}+\frac{1}{1+n}\right)+\frac{\Gamma'(1)}{\Gamma(1)}$$

This result is sometimes written in another form; namely, putting $\frac{\Gamma'(1)}{\Gamma(1)}=-C$ we can write

$$\frac{\Gamma'(z)}{\Gamma(z)}=-\frac{1}{z}+1+\sum_{n=1}^{\infty}\left(-\frac{1}{z+n}+\frac{1}{n}\right)-\sum_{n=1}^{\infty}\left(\frac{1}{n}-\frac{1}{n+1}\right)-C=$$

$$=-C-\frac{1}{z}+\sum_{n=1}^{\infty}\left(\frac{1}{n}-\frac{1}{n+z}\right) \qquad (58)$$

Replacing in (58) z by ζ and integrating with respect to ζ from 1 to z we derive

$$-\ln\Gamma(z)=Cz-C+\ln z-\sum_{n=1}^{\infty}\left(\frac{z-1}{n}-\ln\frac{n+z}{n+1}\right) \qquad (59)$$

Now taking $z=2$ we obtain

$$C=-\ln 2+\sum_{n=1}^{\infty}\left(\frac{1}{n}-\ln\frac{n+2}{n+1}\right)=$$

$$=-\ln 2+\lim_{N\to\infty}\left(\sum_{n=1}^{N}\frac{1}{n}-\ln\frac{3}{2}-\ln\frac{4}{3}-\ldots-\ln\frac{N+2}{N+1}\right)=$$

$$=\lim_{N\to\infty}\left(\sum_{n=1}^{N}\frac{1}{n}-\ln N-\ln\frac{N+2}{N}\right)=$$

$$=\lim_{N\to\infty}\left(\sum_{n=1}^{N}\frac{1}{n}-\ln N\right)-\lim_{N\to\infty}\ln\frac{N+2}{N}=\lim_{N\to\infty}\left(\sum_{n=1}^{N}\frac{1}{n}-\ln N\right)$$

Thus we see that C is nothing but **Euler's constant** (e. g. see [107], Sec. XVII.1). Substituting the expression

$$C=\sum_{n=1}^{\infty}\left(\frac{1}{n}-\ln\frac{n+1}{n}\right)$$

(check it up!) into (59) and raising we arrive at the expansion of the function $\frac{1}{\Gamma(z)}$ as an infinite product of the form

$$\frac{1}{\Gamma(z)}=ze^{Cz}\prod_{n=1}^{\infty}\left(1+\frac{z}{n}\right)e^{-\frac{z}{n}}$$

The logarithmic derivative $\psi_1(z)=\frac{\Gamma'(z)}{\Gamma(z)}$ determined by formula (58) and also its derivatives

$$\psi_2(z)=\sum_{n=0}^{\infty}\frac{1}{(n+z)^2},\ \ldots,\ \psi_p(z)=(-1)^p(p-1)!\sum_{n=0}^{\infty}\frac{1}{(n-z)^p} \quad (60)$$

are used for computing the sums of number series of the form

$$\sum_{n=0}^{\infty}\frac{P(n)}{Q(n)}$$

where P and Q are polynomials such that all the zeros of Q are real and its degree is at least by 2 greater than that of P. To this end, we take the decomposition of the proper rational fraction $\frac{P(z)}{Q(z)}$ into partial fractions (e. g. see [107], Sec. VIII.10) and then apply formulas (60) and (58). We also give here the simple formula

$$\psi_p(z+1)=\psi_p(z)+(-1)^{p-1}(p-1)!\,z^{-p}$$

(let the reader deduce it!) which makes it possible to use for computations connected with the function $\psi_p(x)$ only tables of the values of that function corresponding to an interval of variation of x of length 1.

13. Schwarz-Christoffel Transformation. We shall begin this section with an important application of the argument principle (Sec. 7) to conformal mappings.

Let $f(z)$ be a one-valued analytic function in an open finite simply connected domain (G). Suppose that this function is continuous in the closure of the domain (G) (including its boundary (Γ)) and that the image of (Γ) under the mapping $w=f(z)$ is the boundary (D) of an open finite domain (H) of the w-plane. Then, if the point w makes one circuit along (D) in the positive direction when the point z makes one circuit along (Γ) in the positive di-

rection (it is not supposed originally that the mapping is one-to-one), *the domain* (H) *is a conformal image of* (G), *that is the function* $w=f(z)$ *determines a conformal mapping of the domain* (G) *onto the domain* (H).

In order to prove this assertion we apply formula (41) to the function $f(z)-w_0$ where w_0 is an arbitrary point of (H). This formula shows that there is only one point $z\in(G)$ which is mapped on w_0. Similarly, if the point w_0 is in the exterior of (D) there are no points in the domain (G) which are mapped on w_0. Now applying the results of the first paragraph of Sec. 2.6 we conclude that the mapping is conformal, which is what we set out to prove.

The above proposition makes it possible to limit ourselves to testing the properties of a mapping only on the boundary of the given domain when we are interested whether the mapping is conformal in this domain. It is also readily extended to unbounded domains (by approximating them with finite domains) if the positive direction of describing the boundary of an infinite domain is understood as the one for which the domain always remains on the left.

Let us now consider, in the upper half-plane $\operatorname{Im} z>0$, the function

$$w=C\int_{z_0}^{z}(\zeta-a_1)^{-\alpha_1}(\zeta-a_2)^{-\alpha_2}\ldots(\zeta-a_n)^{-\alpha_n}d\zeta+C_1 \qquad (61)$$

where $C\neq 0$ and C_1 are arbitrary complex constants, $\operatorname{Im} z_0>0$, all a_k and α_k are arbitrary real numbers such that

$$a_1<a_2<\ldots<a_n,\ |\alpha_k|<1 \text{ for all } k, \text{ and } \sum_k \alpha_k=2$$

The integrand can be regarded as a one-valued function in the upper half-plane if we put $(\zeta-a_k)^{\alpha_k-1}=\exp[(\alpha_k-1)\ln(\zeta-a_k)]$ (see Sec. 1.12). Then we can say that function (61) specifies an analytic mapping of the upper half-plane (let us denote it as (G)) of the variable z on a domain (H) of the w-plane; its configuration we are going to determine now.

We have

$$\frac{dw}{dz}=C(z-a_1)^{-\alpha_1}(z-a_2)^{-\alpha_2}\ldots(z-a_n)^{-\alpha_n}$$

and hence when the point z moves along the real axis which is the boundary of the domain (G) the argument of $\frac{dw}{dz}$ remains constant, i.e. the point w moves along a rectilinear segment (see Sec. 1.4) until the point z passes through one of the points a_k. After z has passed the position a_k the magnitude of $\operatorname{Arg}(z-a_k)$

is increased by $-\pi$ and therefore $\operatorname{Arg}\frac{dw}{dz}$ is increased by $\alpha_k\pi$, which means that the point w proceeds to move in a new direction forming an angle $\alpha_k\pi$ with the former one (Fig. 50). Consequently, the x-axis is mapped onto an n-gon with vertices a_1',

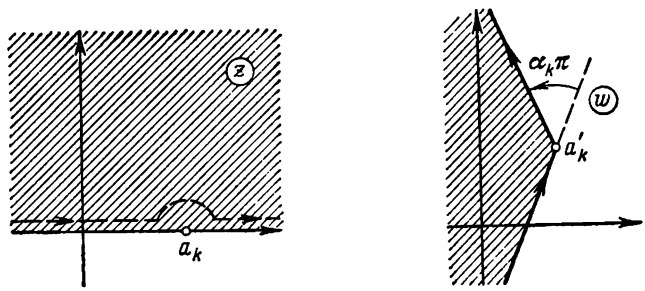


Fig. 50

$a_2', \ldots, a_n'$. (The point $z=\infty$, $\operatorname{Im} z \geqslant 0$, is mapped on one of the points of the line segment $a_n'a_1'$, which can be proved on the basis of the equality $\sum\alpha_k = 2$ but this we leave to the reader.) If the polygon with vertices $a_1', a_2', \ldots, a_n'$ does not have "internal" self-intersections (of the type shown in Fig. 51*c*), the theorem

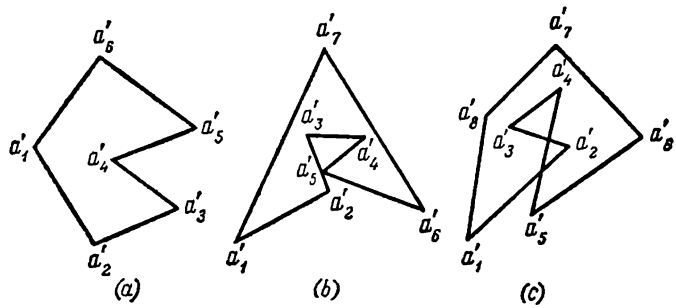


Fig. 51

proved at the beginning of this section implies that *function* (61) *determines a conformal mapping of the upper half-plane on the interior of the n-gon*. (See Fig. 51*a*, *b* and *c*; in the cases *a* and *b* the mapping is possible and in the case *c* it does not exist.) Let the reader show that if $\alpha_k > 0$ for all k or if $n \leqslant 4$ these self-intersections cannot exist. Formula (61) expresses the so-called **Schwarz-Christoffel transformation.**

The converse (which we shall not prove here) is also true: given an n-gon in the w-plane, we can always choose the parameters

in formula (61) in such a way that the upper half-plane of the variable z is mapped on that n-gon (but this choice of the parameters is not simple in practical problems). The following consideration (which is not a rigorous proof of the converse) indicates intuitive reasons for the possibility of choosing the values of the parameters in the appropriate manner: the number of essential real parameters in formula (61) exceeds by 3 the number of real parameters specifying an n-gon and a conformal mapping of the half-plane into itself is also determined by the choice of three real parameters (see Sec. 1.16). In particular, these three (real) degrees of freedom make it possible to set three of the numbers

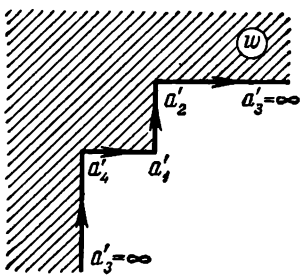


Fig. 52

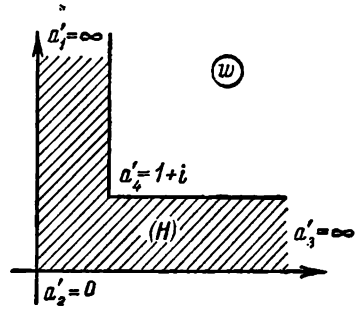


Fig. 53

a_k in an arbitrary way and then determine the others. In applications we often put $a_n=\infty$. To this end, it is advisable to write $C(-a_n)^{\alpha_n}$ instead of C in formula (61) and then pass to the limit for $a_n\to\infty$ which shows that for $a_n=\infty$ the last factor in the integrand should be discarded.

If one of the numbers α_k in formula (61) is not less than unity (but it must not exceed 3), the singularity at $z=a_k$ is not integrable (integral (61) becomes divergent). This means that the corresponding vertex of the "polygon" travels to infinity. When the variable point traverses the contour of the polygon and passes through that vertex it changes the direction of motion which, as before, turns by the angle $\alpha_k\pi$. For instance, let the reader prove that the "quadrilateral" shown in Fig. 52 is specified by the values $\alpha_1=\frac{1}{2}$, $\alpha_2=-\frac{1}{2}$, $\alpha_3=\frac{5}{2}$ and $\alpha_4=-\frac{1}{2}$.

Consider an example. Let it be necessary to construct a conformal mapping of the upper half-plane z onto the "quadrilateral" (H) depicted in Fig. 53. Here we have $\alpha_1=\alpha_3=1$, $\alpha_2=\frac{1}{2}$ and $\alpha_4=-\frac{1}{2}$. Let us put $a_1=-1$, $a_2=0$ and $a_4=\infty$. Then formula

(61) takes the form

$$w = C\int \frac{dz}{(z+1)\sqrt{z}(z-a_3)} = 2C\int \frac{ds}{(s^2+1)(s^2-a_3)} =$$
$$= \frac{2C}{1+a_3}\left(\int \frac{ds}{s^2-a_3} - \int \frac{ds}{s^2+1}\right) =$$
$$= \frac{2C}{1+a_3}\left(\frac{1}{2\sqrt{a_3}}\operatorname{Ln}\frac{s-\sqrt{a_3}}{s+\sqrt{a_3}} - \frac{1}{2i}\operatorname{Ln}\frac{s-i}{s+i}\right) + C_1 \qquad (62)$$

where $z = s^2$, $s = \sqrt{z}$. (When computing the last integral we have used the logarithm which is more convenient for operations on complex quantities than the arctangent.) In further calculations we must take into account that the functions involved are multiple-valued. Since the variable z varies in the upper half-plane, we can take, as the range of s, either the first or the third quadrant. For definiteness, let us take the first quadrant. To specify the branches of the logarithmic function we use the condition $f(a_2) = a_2'$, i.e. $f(0) = 0$, and write

$$\frac{2C}{1+a_3}\left(\frac{1}{2\sqrt{a_3}}\operatorname{Ln}(-1) - \frac{1}{2i}\operatorname{Ln}(-1)\right) + C_1 = 0$$

Expressing C_1 from this relation and substituting it into (62) we obtain

$$w = \frac{C}{1+a_3}\left(\frac{1}{\sqrt{a_3}}\operatorname{Ln}\frac{\sqrt{a_3}-s}{\sqrt{a_3}+s} - \frac{1}{i}\operatorname{Ln}\frac{i-s}{i+s}\right) \qquad (63)$$

and the same condition $w(0) = 0$ indicates that we must take the branches of the logarithm satisfying the equality $\operatorname{Ln} 1 = 0$. It is readily verified that if s ranges over the first quadrant, the fractions in formula (63) which are under the sign of logarithm determine variable points that describe, respectively, the lower and the upper semicircular arcs of unit circle, and consequently for $\operatorname{Im} z > 0$ we must take the principal value of the logarithmic function (why?). Furthermore, for the domain (H) to occupy the proper position it is necessary that the derivative $\frac{dw}{dz}$ be real for real $z > 0$ whence, by the first equality (62), the constant C is also real. Now we must consider the condition $w(\infty) = 1 + i$. Substituting $s = \infty$ into (63) and taking into account the above remark on the ranges of the fractions under the sign of logarithm we obtain

$$1 + i = \frac{C}{1+a_3}\left(\frac{1}{\sqrt{a_3}}(-\pi i) - \frac{1}{i}\pi i\right) \text{ whence } C = -\frac{2}{\pi},\ a_3 = 1$$

Substituting these values into (63) we finally receive

$$w = \frac{1}{\pi}\left(\ln\frac{1+s}{1-s} + i\ln\frac{i+s}{i-s}\right),\quad s = \sqrt{z},\ \operatorname{Im} z > 0$$

Let the reader carefully consider the above argument and particularly take into account that the operations on many-valued functions require specification of the branches of the functions involved because otherwise we can arrive at incorrect results.

14. Elliptic Functions. As a rule, integral (61) is not reducible to elementary functions. In particular, when the half-plane is mapped onto the interior of a rectangle we obtain an *elliptic integral*. As is known (e.g. see [107], Sec. XIII.11), an **elliptic integral** has the form

$$\int R(x, \sqrt{P(x)})\,dx \tag{64}$$

where R is a rational function of its arguments and $P(x)$ is a polynomial of the third or forth degree. When the coefficients of the functions involved are connected by certain relations integral (64) can be expressed in terms of elementary functions (in this case it is referred to as a **quasi-elliptic integral**), but, generally, this is not the case. It can be proved (e.g. see [2]) that integral (64) is always representable as a combination of elementary functions and integrals of the form

$$\int \frac{dx}{\sqrt{(1-x^2)(1-k^2x^2)}}, \quad \int \sqrt{\frac{1-k^2x^2}{1-x^2}}\,dx,$$
$$\int \frac{dx}{(1+n^2x^2)\sqrt{(1-x^2)(1-k^2x^2)}}$$

where k and n are constants. Usually we get rid of arbitrary constants in these indefinite integrals and take concrete antiderivatives in the form of definite integrals with variable upper limits of integration:

$$\int_0^x \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}, \quad \int_0^x \sqrt{\frac{1-k^2t^2}{1-t^2}}\,dt, \quad \int_0^x \frac{dt}{(1+n^2t^2)\sqrt{(1-t^2)(1-k^2t^2)}}$$

Performing in the last integrals the substitution $t = \sin\psi$ we obtain three special functions of the form

$$\int_0^x \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}} = \int_0^\varphi \frac{d\psi}{\sqrt{1-k^2\sin^2\psi}} \equiv F(\varphi, k)$$

$$\int_0^x \sqrt{\frac{1-k^2t^2}{1-t^2}}\,dt = \int_0^\varphi \sqrt{1-k^2\sin^2\psi}\,d\psi \equiv E(\varphi, k) \tag{65}$$

$$\int_0^x \frac{dt}{(1+n^2t^2)\sqrt{(1-t^2)(1-k^2t^2)}} = \int_0^\varphi \frac{d\psi}{(1+n^2\sin^2\psi)\sqrt{1-k^2\sin^2\psi}} \equiv \Pi(\varphi, n, k)$$

where $\sin\varphi = x$. Integrals (65) were called by **Legendre incomplete elliptic integrals of the first, second and third kind** (as functions of φ they are said to be in **Legendre's normal** or **standard form**). It should be noted that the reduction of (64) to integrals (65) may involve very complicated expressions and, besides, integrals of the third kind are inconvenient for calculations and even were not tabulated. Therefore, if an integral of type (64) cannot be reduced to integrals of the first and second kind in a simple manner it is advisable to resort to some approximate or numerical methods.

There are extensive tables of values of $F(\varphi, k)$ and $E(\varphi, k)$ compiled for $0 \leqslant \varphi \leqslant \frac{\pi}{2}$ and $0 \leqslant k \leqslant 1$. The integrands in these integrals are even and periodic (with period 2π) functions of ψ, and, consequently, knowing the values of $F(\varphi, k)$ and $E(\varphi, k)$ for $0 \leqslant \varphi \leqslant \frac{\pi}{2}$ we can easily find their values for any other interval of variation of φ. Similarly, the case $k > 1$ is reduced to the interval $0 \leqslant k \leqslant 1$ by means of the substitution $k \sin\psi = \sin\psi_1$ (check it up for real integrals). The quantity k is called the **modulus** of the integrals and φ is referred to as their **amplitude** (the quantity n is referred to as the **parameter**). In most applications these quantities are real. (Operations on elliptic integrals also involve the quantity $k' = (1-k^2)^{\frac{1}{2}}$ which is called the **complementary modulus**.)

For $\varphi = \frac{\pi}{2}$ we obtain from $F(\varphi, k)$ and $E(\varphi, k)$ the so-called **complete elliptic integrals of the first and second kinds** which are most frequently used:

$$K(k) = F\left(\frac{\pi}{2}, k\right) \quad \text{and} \quad E(k) = E\left(\frac{\pi}{2}, k\right)$$

Let the reader prove that the arc length of an ellipse with eccentricity ε and semi-major axis a is equal to $4aE(\varepsilon)$ (for this purpose it is advisable to take the standard equation of the ellipse and apply the general formula for the arc length of a curve). This fact accounts for the term "elliptic integral".

Let us consider the first integral (65) and suppose that now the independent variable $x = z$ is complex. Comparing it with formula (61) we conclude that the function

$$w = \int_0^z [(1-\zeta^2)(1-k^2\zeta^2)]^{-1/2}\, d\zeta \quad (0 < k < 1) \tag{66}$$

(here we take the single-valued branch of the integrand function which equals unity for $\zeta = 0$) determines a conformal mapping of

the upper half-plane of the variable z on the interior of a rectangle in the w-plane, the points $z=\pm 1$ and $z=\pm\frac{1}{k}$ going into the vertices of the rectangle. For $z=\pm 1$ we obtain $w=\pm K(k)$ and for $z=\pm\frac{1}{k}$ we have

$$w=\pm K(k)+i\int_1^{\frac{1}{k}}[(t^2-1)(1-k^2t^2)]^{-\frac{1}{2}}dt=$$

$$=\pm K(k)+i\int_0^1[(1-\tau^2)(1-k'^2\tau^2)]^{-\frac{1}{2}}d\tau=\pm K(k)+iK(k')$$

(where $t=(1-k'^2\tau^2)^{-\frac{1}{2}}$ and $k'=\sqrt{1-k^2}$; check up the calculations). Thus we have determined the vertices of the rectangle.

The mapping being one-to-one, the inverse of function (66) (denoted as $z=\mathrm{sn}(w, k)$ or, simply, $z=\mathrm{sn}\,w$ when it is unnecessary to emphasize the modulus) generates a conformal mapping of the interior of the above rectangle on the upper half-plane z. Applying the symmetry principle (Sec. 1.16) we can easily show (e.g. see [81]) that the function $\mathrm{sn}\,w$ can be analytically continued from this rectangle to the whole w-plane, which results in a meromorphic function with simple poles at the points $w=iK(k')+2mK(k)+2inK(k')$ where m and n are arbitrary integers. This function satisfies the identity

$$\mathrm{sn}(w+4mK(k)+2inK(k'))\equiv\mathrm{sn}\,w \qquad (m,\ n=0,\ \pm 1,\ \pm 2,\ \ldots)$$

i.e. has two independent periods $4K(k)$ and $2iK(k')$. A function of a complex variable possessing such a property is said to be **doubly periodic.**

The expression $\mathrm{sn}\,w$ is read **"sn"-function** (of w) or **"sinus amplitudinis w"**. (The latter term is accounted for by the fact that if we make in integral (66) the substitution $\zeta=\sin\psi$ similar to the one used above for real integrals we obtain the corresponding complex integral with the variable upper limit of integration φ, $\sin\varphi=z$, which is called *the amplitude of* w and denoted as $\varphi=\mathrm{am}\,w$.) In the theory of elliptic functions we also introduce the functions $\mathrm{cn}\,w=\sqrt{1-\mathrm{sn}^2 w}$ (**"cn"-function**) and $\mathrm{dn}\,w=\sqrt{1-k^2\sin^2 w}$ (**"dn"-function**) satisfying the normalization condition that their values for $w=0$ are equal to unity; they are called, respectively, **"cosinus amplitudinis w"** and **"delta amplitudinis w"**. It can be shown that the functions $\mathrm{cn}\,w$ and $\mathrm{dn}\,w$ are also meromorphic and doubly periodic. The functions $\mathrm{sn}\,w$, $\mathrm{cn}\,w$ and $\mathrm{dn}\,w$ were introduced by Jacobi and are referred to as **Jacobi's elliptic functions.** There are many useful relationships

between these functions. It should be noted that for real w these functions assume real values and are periodic with period $4K(k)$. Their graphs for a concrete value of k are shown in Fig. 54. In particular, we have $\operatorname{sn} x = \sin \varphi(x)$ where the function $\varphi(x)$ is the inverse of the increasing function $x(\varphi) = F(\varphi, k)$.

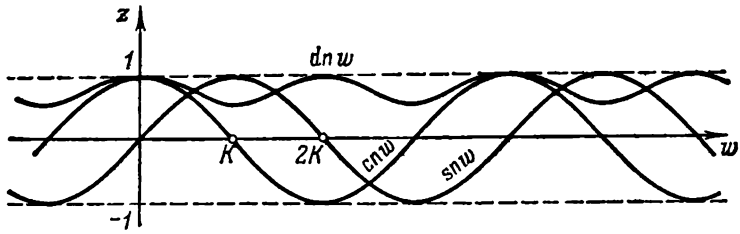


Fig. 54

There are some other types of elliptic functions closely related to those discussed here. The properties of these functions can be found in [2] and in courses concerning special functions.

§ 4. ASYMPTOTIC EXPANSIONS

As is known (e.g. see [107], Sec. XVII.19), asymptotic expansions are widely used in applied mathematics for deriving approximate limiting formulas and for computations. The theory of asymptotic expansions is directly related to the theory of analytic functions (although some of its results deal with purely real relationships). Even in applied problems, it sometimes involves complicated and subtle mathematical considerations and techniques. Here we shall only present some general notions concerning asymptotic expansions; for more detail we refer the reader to [21], [24] and [34].

1. Introduction. We say that a function $f(x)$ possesses the **asymptotic expansion**

$$f(x) \sim a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \ldots + \frac{a_n}{x^n} + \ldots \tag{1}$$

for $x \to \infty$ if the relation

$$f(x) = a_0 + \frac{a_1}{x} + \ldots + \frac{a_n}{x^n} + o\left(\frac{1}{x^n}\right) \quad (\text{for } x \to \infty) \tag{2}$$

holds for every $n = 0, 1, 2, \ldots$. Comparing the expressions given by formula (2) for $n+1$ and for n we see that the remainder $o\left(\frac{1}{x^n}\right)$ can also be written as $O\left(\frac{1}{x^{n+1}}\right)$. It should be noted that the series on the right-hand side of formula (1) is not necessarily

convergent to the function $f(x)$ in the ordinary sense. Moreover, it may be divergent and have no limit function at all. For instance, consider the expansion (e.g. see [107], Sec. XVII.19)

$$\int_x^\infty \frac{e^{x-s}}{s}\,ds \sim \frac{1}{x} - \frac{1!}{x^2} + \frac{2!}{x^3} - \frac{3!}{x^4} + \ldots \quad (x \to \infty) \tag{3}$$

Formula (3) yields an asymptotic expansion of the integral although the series on the right-hand side is divergent for every x. Computing, in succession, the partial sums of this series we find that

for $x=2$ they are equal to

$$0.50;\ 0.25;\ 0.50;\ 0.125;\ 0.875;\ \ldots$$

for $x=5$ we have

$$0.200;\ 0.160;\ 0.176;\ 0.166;\ 0.174;\ 0.166;\ 0.176;\ 0.163;\ 0.183;\ \ldots$$

and for $x=10$

$$0.10000;\ 0.09000;\ 0.09200;\ 0.09140;\ 0.09164;\ 0.09152;\ 0.09159;$$
$$0.09154;\ 0.09158;\ 0.09155;\ 0.09158;\ 0.09154;\ 0.09159;\ \ldots$$

Note that the deduction of formula (3) clearly indicates (check it up!) that for any fixed $x>0$ the remainder changes its sign when we pass from one value of n to the subsequent value. Here we deal with the case which is most convenient for computations because the values of the quantity whose expansion is considered lie between any two subsequent partial sums of the series, which enables us to estimate the upper and lower bounds for these values. For instance, in the above example we have

$$0.25 < f(2) < 0.50;\quad 0.166 < f(5) < 0.174;$$
$$0.09155 < f(10) < 0.09158$$

that is we can approximately put $f(2)=0.4$, $f(5)=0.170$ and $f(10)=0.09156$. We also see that the greater the value of x, the more accurate the result.

If it is difficult to determine directly upper and lower bounds for the quantity we are interested in we sometimes manage to represent the remainder in the form of an integral, which makes it possible to obtain an estimation and thus to have a good judgement on the accuracy of the result. But there are cases when we cannot obtain a precise estimation of the remainder, and then, in practical computations, we usually compute, in succession, partial sums; if we see that their values become closer for a value of x which is not too small we have every reason to take these partial sums as approximate values of the quantity and thus derive a plausible estimation of the error.

Asymptotic expansions of the form

$$f(x) \sim g(x)\left(a_0 + \frac{a_1}{x} + \ldots + \frac{a_n}{x^n} + \ldots\right) \tag{4}$$

and

$$f(x) \sim a_0 + a_1 x^{-p} + a_2 x^{-2p} + \ldots \quad (p > 0)$$

and the like are also used. The sense of these formulas is analogous to that of (1).

If we consider an asymptotic expansion

$$f(z) \sim a_0 + \frac{a_1}{z} + \frac{a_2}{z^2} + \ldots + \frac{a_n}{z^n} + \ldots \tag{5}$$

of a one-valued analytic function $f(z)$ we must stipulate in which ways the point z approaches infinity. If it turns out that formula (5) is valid for any path along which z travels to infinity, it can readily be shown (let the reader try to do it!) that the right-hand side of formula (5) is nothing but Laurent's expansion of the function $f(z)$ at $z = \infty$ (see Sec. 3.6); then equality (5) is not asymptotic but precise. In the theory of asymptotic expansions we usually obtain more significant results when equality (5) is considered on condition that the point z approaches infinity within a given angle, that is when the argument of z satisfies an inequality of the form $\alpha \leqslant \operatorname{Arg} z \leqslant \beta$. (In particular, for $\alpha = \beta = 0$ we obtain an expansion of form (1).) In typical problems we most often set an appropriate inequality $\alpha < \operatorname{Arg} z < \beta$ in such a way that the formula

$$f(z) = a_0 + \frac{a_1}{z} + \ldots + \frac{a_n}{z^n} + o\left(\frac{1}{z^n}\right)$$

is valid when the point z tends to infinity within a closed angular region $\alpha' \leqslant \operatorname{Arg} z \leqslant \beta'$ strictly contained in the above angle, i.e. for $\alpha < \alpha' \leqslant \beta' < \beta$. It should also be noted that in different angles with no rays in common one and the same function $f(z)$ may have different asymptotic expansions (this phenomenon was described by Stokes).

2. Properties. For simplicity, we shall speak about the expansions of form (1) for $x \to \infty$ although the expansions of form (4) possess analogous properties. By the way, the first expansion (4) can be rewritten as $\frac{f(x)}{g(x)} \sim a_0 + \frac{a_1}{x} + \ldots$, and thus is reduced to (1), while the second turns into (1) if we put $x^p = y$.

It should be first of all noted that by far not every function (even a finite one for $x \to \infty$) admits of expansion (1). For instance, there is no such expansion for the functions $f(x) = \sin x$, x^{-p} (where $p > 0$ is a fractional number), $\frac{\ln x}{x}$, etc. On the other hand, it can be easily shown that if an expansion exists its coeffi-

cients are determined uniquely, that is one and the same function $f(x)$ cannot have two different expansions of type (1). Indeed, if, besides (1), we have the formula $f(x) \sim a_0' + \frac{a_1'}{x} + \ldots$, and $n \geqslant 0$ is the least number for which $a_n \neq a_n'$, we can write

$$f(x) = a_0' + \frac{a_1'}{x} + \ldots + \frac{a_n'}{x^n} + o\left(\frac{1}{x^n}\right) = a_0 + \frac{a_1}{x} + \\ + \ldots + \frac{a_{n-1}}{x^{n-1}} + \frac{a_n'}{x^n} + o\left(\frac{1}{x^n}\right)$$

Now, performing the subtraction we arrive at a contradiction (why?). When we stipulated in Sec. 1 that two angular regions, in which an analytic function may possess different asymptotic expansions, must have no rays in common we took into account the property proved here.

At the same time, two different functions can have the same expansion (1). In fact, every function that tends to zero for $x \to \infty$ faster than any power of x (for instance, such is the function e^{-x}) has an asymptotic expansion whose all coefficients are equal to zero (why?). Hence, if we add a function of that type to the left-hand side of (1), the right-hand side remains unchanged. Consequently, a given asymptotic expansion determines the corresponding function to within a summand of the above type. But usually this arbitrariness is inessential in practical problems.

Let us proceed to study operations on expansions of type (1). On the basis of formula (2) it can be readily shown that these expansions can always be added together term-by-term and multiplied by a number (if we multiply (1) by a function we should pass to the first expansion (4)). It can also be proved that expansions of form (1) can be multiplied by one another according to the rule of multiplication of polynomials and also divided by means of the ordinary process of long division or by the method of undetermined coefficients (e.g. see [107], Sec. XVII.12) if the coefficient a_0 in the divisor is nonzero. We can also obtain an asymptotic expansion of a composite function $F(f(x))$ by substituting one series into another if the function $F(y)$ can be expanded into Taylor's series in powers of $y - a_0$; the same applies to more complicated expressions of the form $F(f_1(x), f_2(x), \ldots, f_k(x))$.

Asymptotic expansion (1) can be integrated termwise, which results in

$$\int_{x_0}^{x} f(s)\,ds - a_0 x - a_1 \ln x \sim b_0 - \frac{a_2}{x} - \frac{a_3}{2x^2} - \frac{a_4}{3x^3} - \ldots \qquad (6)$$

where b_0 is a constant dependent on $x_0 \leqslant \infty$. To prove (6) we replace x in formula (2) by s, transfer the first two terms

from the right-hand side to the left-hand side and integrate the result with respect to s from x to ∞ taking into account that $\int_x^\infty o\left(\frac{1}{s^p}\right) ds = o\left(\frac{1}{x^{p-1}}\right)$ $(p > 1,\ x \to \infty)$, which yields

$$\int_x^\infty \left[f(s) - a_0 - \frac{a_1}{s}\right] ds \sim \frac{a_2}{x} + \frac{a_3}{2x^2} + \frac{a_4}{3x^3} + \ldots$$

Changing the signs in both members and adding the term $\int_{x_0}^\infty \left[f(s) - a_0 - \frac{a_1}{s}\right] ds = \int_{x_0}^\infty O\left(\frac{1}{s^2}\right) ds = \text{const}$, we arrive at (6) (check it up!).

The property which has been proved implies that equality (1) can also be differentiated termwise provided the function $f'(x)$ has an asymptotic expansion, that is

$$f'(x) \sim -\frac{a_1}{x^2} - \frac{2a_2}{x^3} - \frac{3a_3}{x^4} - \ldots \tag{7}$$

(Let the reader prove this assertion by taking the asymptotic expansion of $f'(x)$ and showing, with the aid of integration, that this expansion must be of form (7).)

If the above condition that $f'(x)$ possesses an asymptotic expansion does not hold it is possible to construct some artificial examples in which term-by-term differentiation of an asymptotic expansion is illegitimate. But in practical problems we do not encounter such cases. This is partly due to the fact that, as it can be proved, an expansion of form (5) valid for the interior of an angle $\alpha < \operatorname{Arg} z < \beta$ (in the sense indicated at the end of Sec. 1) can always be differentiated termwise, and the derived formula has the same asymptotic sense as the original one (in practical problems we usually deal with expansions which are valid for an angle $|\operatorname{Arg} z| < \varepsilon$ provided that they hold for $x \to \infty$).

We note that in more complicated problems it is sometimes impossible to obtain expansion (1) with an infinite asymptotic series and then we can try to derive a formula of type (2) containing a fixed number n of terms. These formulas are extremely useful for large n but can also be of considerable importance even in the case $n = 0$.

We also consider asymptotic expansions for functions of several variables, but it should be taken into account that these expansions may differ for different relationships between the arguments. For example, take the function

$$f(x, y) = \frac{x}{y} + \sin\frac{1}{x+y} + \sin\frac{1}{x^2} \tag{8}$$

Let us consider it for $x \to \infty$ and $y \to \infty$. Then, if x and y are of the same order, that is if the ratio $k = \frac{y}{x}$ is contained between two positive constants and, in particular, can itself be constant, the expansion is of the form

$$f(x, kx) = \frac{1}{k} + \frac{1}{(1+k)x} + \frac{1}{x^2} - \frac{1}{3!(1+k)^3 x^3} + \dots$$

If y is of the second order relative to x, i. e. $y = kx^2$, we have

$$f(x, kx^2) = \frac{1}{kx} + \frac{k+1}{kx^2} - \frac{1}{k^2 x^3} + \dots$$

In (8) we can put $y = \infty$ which results in

$$f(x, \infty) = \frac{1}{x^2} - \frac{1}{3!\,x^6} + \frac{1}{5!\,x^{10}} - \dots$$

But, on the other hand, it is impossible to put $x = \infty$ in (8).

The relationships between the arguments must always be carefully stipulated because otherwise we can arrive at incorrect results.

3. The Fourier-Type Integral. Let us consider the integral

$$I(\nu) = \int_a^b f(x)\, e^{i\nu x}\, dx \tag{9}$$

where the function $f(x)$ possesses continuous derivatives of all orders on the finite interval $a \leqslant x \leqslant b$. These integrals (with finite or infinite limits) appear in the theory of *Fourier's integral transformation* (e. g. see [107], § XVII.5) and are referred to as **Fourier-type integrals**. We can easily obtain the asymptotic expansion of form (1), for integral (9), with respect to the variable ν for real $\nu \to \pm\infty$. To this end, we integrate (9) by parts several times, for instance, twice, by differentiating the function $f(x)$:

$$I(\nu) = f(x)\frac{e^{i\nu x}}{i\nu}\Big|_a^b - \frac{1}{i\nu}\int_a^b f'(x)\, e^{i\nu x}\, dx =$$

$$= \left(f(x)\frac{e^{i\nu x}}{i\nu} - f'(x)\frac{e^{i\nu x}}{(i\nu)^2}\right)\Big|_a^b + \frac{1}{(i\nu)^2}\int_a^b f''(x) e^{i\nu x}\, dx \tag{10}$$

(in the first integration we have put $f(x) = u$, $e^{i\nu x}\, dx = dv$ and in the second $f'(x) = u$, $e^{i\nu x}\, dx = dv$). First of all we see that $I(\nu) \to 0$ for $\nu \to \pm\infty$. The last integral in (10) is of the same type as (9) and hence it tends to zero for $\nu \to \pm\infty$. Therefore, proceeding to integrate by parts, we arrive at the asymptotic expansion

$$I(\nu) \sim e^{i\nu a} \sum_{n=1}^{\infty} \frac{i^n}{\nu^n} f^{(n-1)}(a) - e^{i\nu b} \sum_{n=1}^{\infty} \frac{i}{\nu^n} f^{(n-1)}(b) \tag{11}$$

(check it up!).

Formula (11) also remains true for $a=-\infty$ (the first sum in (11) turns into zero in this case) and for $b=\infty$ (in the latter case the second sum turns into zero) if the corresponding integrals (9) are absolutely convergent. Indeed, for a wide class of functions $\varphi(x)$ the convergence of the integral $\int_a^\infty |\varphi(x)|\,dx$ implies that $\varphi(\infty)=0$ and $\int_a^\infty |\varphi'(x)|\,dx<\infty$. The latter condition makes it possible to pass to the subsequent derivative, etc. This condition is fulfilled if, for instance, the function $\varphi(x)$ has a finite number of intervals of monotonicity because on each of them the relation $\int_\alpha^\beta |\varphi'(x)|\,dx=|\varphi(\beta)-\varphi(\alpha)|$ holds. But on the other hand, there are some artificial counterexamples in which what has been said is violated. For instance, for the function $\varphi(x)=x\exp\left(e^x\ln\frac{2+\sin x}{3}\right)$ the integral $\int_0^\infty |\varphi(x)|\,dx$ is convergent but the condition $\varphi(\infty)=0$ is not fulfilled (this is due to the fact that the graph of the function has "very acute peaks" which approach infinity as $x\to\infty$). The function $\varphi(x)=\frac{1}{x^2}\sin x^2$ is an example in which the condition $\int_a^\infty |\varphi(x)|\,dx<\infty$ does not imply $\int_a^\infty |\varphi'(x)|\,dx<\infty$ because the frequency with which the values of the function oscillate tends to infinity; let the reader examine these instances. But here and henceforward we disregard such artificial cases and do not make stipulations of this kind. More precise formulation including all the necessary requirements can be found in special monographs.

From the above considerations we can draw some conclusions concerning the behaviour of *Fourier's transform* of an absolutely integrable function $f(x)$ defined on the interval $-\infty<x<\infty$. Suppose that there is an integer $m\geqslant 0$ such that all the derivatives of the function of order less than m are continuous while the derivative $f^{(m)}(x)$ has at least one discontinuity of the first kind (jump discontinuity). Then, breaking up the interval of integration of the integral

$$\hat{f}(k)=\frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)\,e^{-ikx}\,dx$$

($\hat{f}$ is the **Fourier transform** of f) into parts corresponding to the intervals of continuity of $f^{(m)}(x)$ and integrating by parts $m+1$

times we conclude, as described above, that $\hat{f}(k)$ tends to zero for $k \to \pm\infty$ and is of the order of $|k|^{-(m+1)}$ (e. g. cf. example at the end of Sec. XVII.32 in [107]). Similarly, if all the derivatives of the function $f(x)$ are continuous for $-\infty < x < \infty$, then $\hat{f}(k) \to 0$ for $k \to \pm\infty$ faster than any negative power of k.

Let us also consider the case when the limits of integration in integral (8) are finite but the integrand approaches infinity at one of the end-points of the interval of integration. For instance, take the integral

$$I(\nu) = \int_a^b (x-a)^{\alpha-1} f(x)\, e^{i\nu x}\, dx \tag{12}$$

where $0 < \alpha < 1$ (this condition guarantees the convergence of the integral) and the function $f(x)$ possesses the same properties as in integral (9). We also suppose that $f(a) \neq 0$ because, if otherwise, we can perform several integrations by parts and thus reduce the problem to this special case. We shall limit ourselves to computing the principal term of the asymptotic expansion of integral (12) for $\nu \to \infty$. To this end, we first represent $I(\nu)$ in the form

$$I(\nu) = \int_a^b (x-a)^{\alpha-1} f(a)\, e^{i\nu x}\, dx + \int_a^b (x-a)^{\alpha} \frac{f(x)-f(a)}{x-a}\, e^{i\nu x}\, dx$$

The integrand in the second integral being continuous, integration by parts indicates that this integral is of the order of $\frac{1}{\nu}$. In the first integral we make the substitution $x = a + \frac{t}{\nu}$, which results in

$$f(a) \int_0^{\nu(b-a)} t^{\alpha-1} \nu^{1-\alpha} e^{ia\nu} e^{it} \frac{dt}{\nu} = f(a)\, e^{ia\nu} \nu^{-\alpha} \int_0^{\nu(b-a)} t^{\alpha-1} e^{it}\, dt$$

The latter integral tends, for $\nu \to \infty$, to the integral $\int_0^{\infty} t^{\alpha-1} e^{it}\, dt$ which, as can be easily shown, is equal to $e^{\frac{i\pi\alpha}{2}} \Gamma(\alpha)$. To obtain this value of the integral we consider the auxiliary integral

$$J = \oint_{L_{R\rho}} \exp\left[iz + (\alpha-1)\ln z\right] dz$$

taken over the closed contour in the complex z-plane ($z = t + is$) depicted in Fig. 55. By Cauchy's theorem (see Sec. 2.2) we conclude that $J = 0$. For $\rho \to 0$ and $R \to \infty$ the integrals along the corresponding circular arcs tend to zero, which is implied by in-

equality (2.3) and Jordan's lemma (Sec. 3.4). Therefore, passing to the limit, we obtain

$$\int_0^\infty \exp[it+(\alpha-1)\ln t]\,dt = -\int_\infty^0 \exp[i\cdot is+(\alpha-1)\ln is]\,i\,ds =$$

$$= i\int_0^\infty e^{-s} e^{\frac{i(\alpha-1)\pi}{2}} s^{\alpha-1} ds = e^{\frac{i\pi\alpha}{2}}\Gamma(\alpha)$$

Now, summarizing, we can write

$$I(\nu) = f(a)\,e^{\frac{i\pi\alpha}{2}}\Gamma(\alpha)\,e^{ia\nu}\nu^{-\alpha} + o(\nu^{-\alpha}) \quad (\nu\to\infty) \tag{13}$$

(let the reader carefully study these considerations). It can also be readily shown that instead of $o(\nu^{-\alpha})$ we can write $O(\nu^{-1})$ here. On further terms of the expansion see [21].

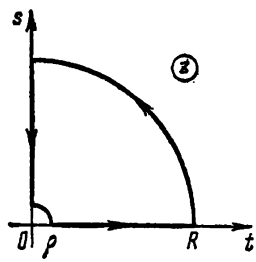


Fig. 55

These results can be applied to a more general integral of the form

$$I_1(\nu) = \int_a^b f(x)\,e^{i\nu\varphi(x)}\,dx$$

with continuous functions f, φ and φ'. If $\varphi'(x) > 0$ $(a \leqslant x \leqslant b)$ we can make the change of variable $\varphi(x) = y$, which yields

$$I_1(\nu) = \int_{\varphi(a)}^{\varphi(b)} f(\varphi^-(y))\,\varphi^{-\prime}(y)\,e^{i\nu y}\,dy \tag{14}$$

where φ^- designates the inverse function of φ (i. e. if $y = \varphi(x)$ then $x = \varphi^-(y)$). Thus we have obtained an integral of type (9) for which expansion (11) is valid. Limiting ourselves to the dominant term we can write, for $\nu \to \pm\infty$, the relation

$$I_1(\nu) = \frac{i}{\nu}\left[e^{i\nu y} f(\varphi^-(y))\varphi^{-\prime}(y)\right]_{\varphi(b)}^{\varphi(a)} + O\left(\frac{1}{\nu^2}\right) = \frac{i}{\nu}\left[e^{i\nu\varphi(x)}\frac{f(x)}{\varphi'(x)}\right]_b^a + O\left(\frac{1}{\nu^2}\right)$$

(check it up!).

The same formula is obtained if $\varphi'(x) < 0$ $(a \leqslant x \leqslant b)$. But if the function $\varphi(x)$ has, on the interval $a \leqslant x \leqslant b$, at least one stationary point the asymptotic formula is considerably changed, which is illustrated below. It turns out that the principal (dominant) term of the asymptotic expansion of $I_1(\nu)$ is specified by the behaviour of the function at its stationary points. First of all we note that we can pass to the case when there is only one stationary point of the function $\varphi(x)$, on the interval under consideration, coinciding with one of its end-points because we can always break up the original interval into parts in an appropriate manner. Let

us again denote the end-points of an interval of that kind by a and b and suppose, for definiteness, that $\varphi'(a)=0$, $\varphi'(x)>0$ for $a<x\leqslant b$ and $\varphi''(a)>0$. All the other possible cases are treated similarly. Making the change of variable $y=\varphi(x)$ in the vicinity of the point $x=a$ we can write, to within infinitesimals of higher order, the expression $y=\varphi(a)+\frac{\varphi''(a)}{2}(x-a)^2$ whence

$$\varphi^{-}(y)=a+\sqrt{\frac{2}{\varphi''(a)}(y-\varphi(a))},\quad \varphi^{-1}(y)=[2\varphi''(a)(y-\varphi(a))]^{-\frac{1}{2}}$$

Therefore integral (14) belongs to type (12), and formula (13) yields

$$I_1(\nu)=\frac{1}{\sqrt{2\varphi''(a)}}f(\varphi^{-}(\varphi(a)))e^{\frac{i\pi}{4}}\Gamma\left(\frac{1}{2}\right)e^{i\varphi(a)\nu}\nu^{-\frac{1}{2}}+o\left(\nu^{-\frac{1}{2}}\right)=$$

$$=\frac{\sqrt{\pi}}{2\sqrt{\varphi''(a)}}f(a)(1+i)e^{i\varphi(a)\nu}\frac{1}{\sqrt{\nu}}+o\left(\frac{1}{\sqrt{\nu}}\right)$$

Here it is also possible to prove that $o\left(\frac{1}{\sqrt{\nu}}\right)$ can in fact be replaced by $O\left(\frac{1}{\nu}\right)$.

4. Integral of Type $I(\nu)=\int_a^b f(x)e^{\nu\varphi(x)}dx$. Now let us proceed to investigate the integral

$$I(\nu)=\int_a^b f(x)e^{\nu\varphi(x)}dx \quad (\nu\to\infty) \tag{15}$$

with real continuous functions f and φ. The behaviour of $I(\nu)$ for large ν essentially depends on the values of x for which $\varphi(x)$ attains its greatest value on the interval $a\leqslant x\leqslant b$ (why?). As in Sec. 3, we can suppose that there is only one such value of x coinciding with an end-point of the interval of integration. For definiteness, let $\varphi(a)>\varphi(x)$ $(a<x\leqslant b)$. We also suppose that

$$\varphi(x)-\varphi(a)\sim -k(x-a)^{\alpha} \qquad (k,\ \alpha>0,\ x\to a+0) \tag{16}$$

and $f(a)\neq 0$. Like in Sec. 2, replacing in integral (15) $f(x)$ by $f(a)$ and $\varphi(x)$ by its expression given by formula (16) we obtain, for $\nu\to\infty$, the relation

$$I(\nu)\sim\int_a^b f(a)e^{\nu[\varphi(a)-k(x-a)^{\alpha}]}dx=$$

$$=f(a)e^{\nu\varphi(a)}\int_0^{k\nu(b-a)^{\alpha}}e^{-y}(k\nu)^{-\frac{1}{\alpha}}\alpha^{-1}y^{\frac{1}{\alpha}-1}dy\sim$$

$$\sim\frac{f(a)}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)e^{\nu\varphi(a)}(k\nu)^{-\frac{1}{\alpha}} \tag{17}$$

(we have made the change of variable $k\nu(x-a)^\alpha = y$). The justification of these transformations lies in the fact that, given any $\varepsilon > 0$, we have

$$\left| \int_{a+\varepsilon}^{b} f(x) e^{\nu\varphi(x)}\, dx \right| \leqslant \int_{a}^{b} |f(x)|\, dx \cdot \exp\left[\nu \max_{a+\varepsilon \leqslant x \leqslant b} \varphi(x)\right]$$

which indicates that this expression is considerably smaller than the right-hand side of (17), and therefore the dominant contribution to integral (15) is made by integration over the interval $a \leqslant x \leqslant a-\varepsilon$. A more rigorous mathematical proof can be found in special monographs. If $\varphi(x)$ attains its maximum value at the point $x=b$, and $\varphi(x)-\varphi(b) \sim -k(b-x)^\alpha$ $(x \to b-0)$, we should replace a by b in the right-hand side of (17). If the maximum value is achieved at an interior point c of the interval of integration and $\varphi(x)-\varphi(c) \sim -k|x-c|^\alpha$ $(x \to c)$, then breaking up the interval into two parts and considering the corresponding integrals we see that in the right-hand side of (17) (where a must be replaced by c) there appears an additional factor equal to 2.

As an example, consider the asymptotic behaviour of the gamma function

$$\Gamma(p) = \int_0^\infty e^{-x} x^{p-1}\, dx = \int_0^\infty \exp\left\{p\left[-\frac{x}{p} + \frac{p-1}{p} \ln x\right]\right\} dx$$

for $p \to \infty$. The expression in the square brackets assumes its maximum value at the point $x=p-1$ (why?), and therefore we make the change of variable $x=(p-1)t$ so that the point of maximum becomes fixed. Thus we obtain

$$\Gamma(p) = (p-1)^p \int_0^\infty \exp\{(p-1)[-t+\ln t]\}\, dt$$

Putting $\varphi(t) = -t \ln t$ we see that $\varphi'(1)=0$ and $\varphi''(1)=-1$, that is $\varphi(t)-\varphi(1) \sim -\frac{1}{2}(t-1)^2$ $(t \to 1)$, and therefore, after formula (17) has been appropriately changed, we receive

$$\Gamma(p) \sim (p-1)^p \cdot 2 \cdot \frac{1}{2} \cdot \Gamma\left(\frac{1}{2}\right) e^{(p-1)(-1)} \left[\frac{1}{2}(p-1)\right]^{-\frac{1}{2}} \sim \sqrt{2\pi} p^{p-\frac{1}{2}} e^{-p}$$

(It should be noted that for $p \to \infty$ the expression $(p-1)^p$ is equivalent to $\frac{p^p}{e}$ but not to p^p as one may think.) This important formula implies that

$$n! = \Gamma(n+1) \sim \sqrt{2\pi}(n+1)^{n+\frac{1}{2}} e^{-n+1} \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \qquad (n \to \infty) \tag{18}$$

Formula (18) is called **Stirling's formula** after J. Stirling (1692-1770), a Scotch mathematician who obtained in 1730 the asymptotic expansion for $\ln(n!)$ which directly implies formula (18). (This was the first application of a series converging asymptotically but not in the ordinary sense.) In 1812 P. S. Laplace derived a more precise formula

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \left(1 + \frac{1}{12n} + \frac{1}{288n^2} - \dots\right)$$

which in particular implies that formula (18) yields a relative error of the order of $\frac{1}{12n}$ while the introduction of the factor $1 + \frac{1}{12n}$ into the right-hand side reduces the error to $\frac{1}{288n^2}$.

In some cases it is possible to apply the above method to integrals of a more general form than (15) in which the functions f and φ depend not only on x but also on ν. Let us consider this procedure by taking a concrete example of the integral

$$K_\nu(a) = \frac{1}{2} \int_{-\infty}^{\infty} \exp(\nu x - a \cosh x)\, dx \quad (0 < a = \text{const},\ \nu \to \infty)$$

playing an important role in the theory of Bessel's functions. Here the exponent attains its maximum for the value $x = \sinh^{-1} \frac{\nu}{a}$ which depends on ν, and therefore it is convenient to make the change of variable $x = \sinh^{-1} \frac{\nu}{2} + t$. This results (check it up!) in

$$K_\nu(a) = \frac{1}{2} \left(\frac{\nu + \sqrt{\nu^2 + a^2}}{a}\right)^\nu \int_{-\infty}^{\infty} \exp\left(\frac{-a^2 \cosh t}{\nu + \sqrt{\nu^2 + a^2}}\right) e^{\nu(t - e^t)}\, dt \quad (19)$$

We have obtained an integral of form (15) in which the function of the type of f also depends on ν. But the behaviour of this function for large ν makes it possible to apply the above technique to integral (19) (why?). This yields the following result for $\nu \to \infty$:

$$K_\nu(a) \sim \frac{1}{2} \left(\frac{\nu + \sqrt{\nu^2 + a^2}}{a}\right)^\nu \cdot 2 \cdot \frac{1}{2} \Gamma\left(\frac{1}{2}\right) e^{-\nu} \left(\frac{\nu}{2}\right)^{-\frac{1}{2}} \sim$$

$$\sim \sqrt{\pi}\, (ea)^{-\nu} (2\nu)^{\nu - \frac{1}{2}}$$

Now let us proceed to determine the subsequent terms of asymptotic expansion (17). For simplicity's sake we shall suppose that $a = 0$, $\varphi(a) = 0$, $\alpha = 1$ and $k = 1$ (the general case is reduced to this one by means of simple changes of variables). Let the expan-

sions of f and φ in the vicinity of the point $x=0$ be of the form

$$f(x)=\sum_{k=1}^{\infty} a_k x^{\alpha_k} \quad (\alpha_1<\alpha_2<\alpha_3<\ldots\longrightarrow\infty,\ \alpha_1\neq 0)$$

$$\varphi(x)=-x+\sum_{k=1}^{\infty} b_k x^{\beta_k} \quad (1<\beta_1<\beta_2<\beta_3<\ldots\longrightarrow\infty)$$

Performing in integral (15) the substitution $vx=s$ we obtain

$$I(v)=\int_0^{vb} v^{-(\alpha_1+1)} s^{\alpha_1} e^{-s}\left[a_1+\sum_{k=2}^{\infty} a_k v^{-(\alpha_k-\alpha_1)} s^{\alpha_k-\alpha_1}\right]\times$$

$$\times\exp\sum_{k=1}^{\infty} b_k v^{-(\beta_k-1)} s^{\beta_k}\,ds \tag{20}$$

Now expanding the exponential function into Maclaurin's series, opening the brackets (including the square brackets in (20)) and combining similar terms we arrive at the expression

$$I(v)=\int_0^{vb} v^{-(\alpha_1+1)} s^{\alpha_1} e^{-s}\left(a_1+\sum_{k=2}^{\infty} c_k v^{-\gamma_k} s^{\delta_k}\right)ds \tag{21}$$

where

$$0<\gamma_1\leqslant\gamma_2\leqslant\gamma_3\leqslant\ldots\longrightarrow\infty,\ \delta_k\geqslant\gamma_k$$

To obtain expansion (21) in explicit form we can set a concrete value of the exponent n of the power of v specifying the accuracy of the asymptotic expansion which is to be constructed and discard all the terms with higher powers appearing in the calculations; e.g. see examples in [107], Sec. XVII.12. We can also take the product of the expression in the square brackets by the exponential function in (20), which is equal to $f\left(\frac{s}{v}\right)\left(\frac{s}{v}\right)^{-\alpha_1}\times$ $\times\exp\left\{\varphi\left(\frac{s}{v}\right)+\frac{s}{v}\right\}$, and make the change of variable $s=t^k$ in it. After this change all the exponents in powers of t become integral, and it is possible to apply the expansion into Maclaurin's series in powers of t.

Integral (21) can be written in the form of a sum of integrals in which the upper limits of integration are replaced by ∞, which does not affect the validity of the asymptotic expansion (why?). This finally yields the formula

$$I(v)\sim a_1\Gamma(\alpha_1+1)\,v^{-(\alpha_1+1)}+\sum_{k=2}^{r} c_k\Gamma(\alpha_1+\delta_k+1)\,v^{-(\alpha_1+\gamma_k+1)}$$

5. Saddle-Point Method. If we suppose that the function $f(x)$ considered in Secs. 3 and 4 can assume complex values the vali-

dity of the asymptotic expansions obtained there is not affected. But if the function $\varphi(x)$ assumes complex values the situation becomes more complicated because a point x_0 at which the expression $|e^{\nu\varphi(x)}|$ (i.e. $e^{\nu \mathrm{Re}\varphi(x)}$) attains its maximum value is not necessarily a stationary point for $\mathrm{Im}\,\varphi(x)$, which means that the values of $e^{\nu\varphi(x)}$ may rapidly oscillate in the vicinity of x_0 in such a way that they mutually cancel out, and then the major contribution to the asymptotic expansion of the integral does not come from these values. Therefore, if we simply introduce a parametric equation $z=z(t)$ of the contour (L) over which the complex integral

$$I(\nu)=\int_{(L)} f(z)\,e^{\nu\varphi(z)}\,dz \qquad (22)$$

is taken, pass to the new variable t and then directly apply the results of Secs. 3 and 4, the expression thus obtained may differ from the sought-for asymptotic expansion of the integral. But if f and φ are single-valued analytic functions we can appropriately deform the contour of integration (L) without affecting the value of integral (22) (see Sec. 2.2) and then construct the required asymptotic expansion.

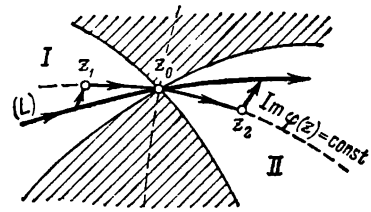


Fig. 56

Suppose that, for the given function $\varphi(z)=u(x, y)+iv(x, y)$, it is possible to choose the contour (L) in such a way that it passes through a point z_0 at which $\varphi'(z_0)=0$ and $|e^{\varphi(z)}|$ attains its maximum value (on the contour (L)). In Fig. 56 we see the neighbourhood of the point z_0 depicted for the case $\varphi''(z_0)\neq 0$. The domain in which $|e^{\varphi(z)}|>|e^{\varphi(z_0)}|$, i. e. $\mathrm{Re}\,\varphi(z)>\mathrm{Re}\,\varphi(z_0)$, is shaded, and the curve $\mathrm{Im}\,\varphi(z)=\mathrm{Im}\,\varphi(z_0)$ is given in dotted line. To construct this domain we must represent $\varphi(z)$, in the vicinity of z_0, in the form $\varphi(z)=\varphi(z_0)+\frac{\varphi''(z_0)}{2!}(z-z_0)^2+\ldots$ and put $z=x+iy$ and $\varphi''(z_0)=a+ib$. Then the above relation can be written, to within the terms of higher order, as $\mathrm{Re}\{(a+ib)\times[(x-x_0)+i(y-y_0)]^2\}>0$, that is in the form $a(x-x_0)^2-2b(x-x_0)(y-y_0)-a(y-y_0)^2>0$. (We suggest that the reader realize a similar scheme for the case $\varphi''(z_0)=0$.) Let the contour (L) be chosen in such a way that it passes through the point z_0 (which is a *hyperbolic point* or *saddle-point* of the surface $u=u(x, y)=\mathrm{Re}\,\varphi(z)$; e. g. see [107], Secs. XII.9 and XV.7) from one "hollow" *I* (i. e. one of the domains in which $\mathrm{Re}\,\varphi(z)<\mathrm{Re}\,\varphi(z_0)$) into another "hollow" *II* in the vicinity of z_0. (This accounts for

the term **"saddle-point method"** applied to the method of deriving an asymptotic formula on the basis of a deformation of the original contour of integration into another contour possessing the above property.)

Let us deform a small portion of the contour (L) (shown in Fig. 56 in heavy line) in such a way that after the deformation a part of the contour passes through the point z_0 along the line $\operatorname{Im}\varphi(z)=\operatorname{Im}\varphi(z_0)$. According to Sec. 1.4, this line is directed at each point (if we move from z_0) along the vector $-\operatorname{grad}\operatorname{Re}\varphi(z)$, and therefore along the vector $-\operatorname{grad}|e^{\varphi(z)}|$. This accounts for the term **"the method of steepest descent"** applied to a deformation of this kind. If the arc $z_1z_0z_2$ (see Fig. 56) is represented in parametric form $z=z(t)$ $(a\leqslant t\leqslant b)$, the corresponding part of integral (22) can be written as

$$\int_a^b f(z(t))\,z'(t)\,e^{\nu[\operatorname{Re}\varphi(z(t))+i\operatorname{Im}\varphi(z(t))]}\,dt=e^{i\nu\operatorname{Im}\varphi(z_0)}\int_a^b f_1(t)\,e^{\nu\varphi_1(t)}\,dt \tag{23}$$

where $f_1(t)=f(z(t))\,z'(t)$ and $\varphi_1(t)=\operatorname{Re}\varphi(z(t))$. To the last integral we can apply the technique used for real $\nu\to\infty$ in Sec. 4 and in some cases even obtain an infinite asymptotic series. The expansion thus obtained serves as asymptotic expansion of integral (22) as well because, like in Sec. 4, we can show that the integral over the contour (L) with the arc $z_1z_0z_2$ cut out from it is considerably smaller than each term of the expansion for $\nu\to\infty$.

We shall limit ourselves to investigating the case $\varphi''(z_0)\neq 0$, $f(z_0)\neq 0$ and finding the first term of the expansion of integral (22) for $\nu\to\infty$. In the vicinity of the point z_0 we have $\varphi(z)=\varphi(z_0)+\frac{\varphi''(z_0)}{2}(z-z_0)^2+\ldots$, and along the arc $z_1z_0z_2$ only the expression $\operatorname{Re}\varphi(z)$ varies (it decreases as the point z moves away from z_0), and therefore we can introduce the parameter t according to the formula $\varphi(z)=\varphi(z_0)-t^2$ whence $z=z_0+$ $+\sqrt{\frac{-2}{\varphi''(z_0)}}\,t+\ldots$ where we have taken the branch of the square root for which the point z passes from region I to region II when t increases and passes through the value $t=0$. Applying formula (17) (which should be appropriately changed) to the right-hand side of (23) we obtain

$$\begin{aligned}I(\nu)&\sim e^{i\nu\operatorname{Im}\varphi(z_0)}\cdot 2\frac{f_1(0)}{2}\Gamma\left(\frac{1}{2}\right)e^{\nu\varphi_1(0)}\left(-\frac{\varphi_1''(0)}{2}\nu\right)^{-\frac{1}{2}}=\\&=e^{i\nu\operatorname{Im}\varphi(z_0)}f(z_0)\sqrt{\frac{-2}{\varphi''(z_0)}}\sqrt{\pi}e^{\nu\operatorname{Re}\varphi(z_0)}\nu^{-\frac{1}{2}}=\\&=f(z_0)\sqrt{\frac{-2\pi}{\varphi''(z_0)\,\nu}}e^{\nu\varphi(z_0)}\end{aligned} \tag{24}$$

It can be proved that this asymptotic equality can be written more precisely in the form

$$I(\nu)=\left[f(z_0)+O\left(\frac{1}{\nu}\right)\right]\sqrt{\frac{-2\pi}{\varphi''(z_0)\,\nu}}\,e^{\nu\varphi(z_0)}$$

which indicates the order of the error. This formula also holds when $f(z_0)=0$.

Let us consider an example. Take the integral

$$P_n(\mu)=\frac{1}{2^{n+1}\pi i}\oint\limits_{(L)}\frac{(z^2-1)^n}{(z-\mu)^{n+1}}\,dz \tag{25}$$

where (L) is a closed contour (containing the point $z=\mu$ inside it) which is traversed in the positive direction. It can be shown that the functions P_n are nothing but **Legendre's polynomials** (e. g. see the definition of Legendre's polynomials in [107], Sec. XVII.20). Let us regard μ as a real constant satisfying the condition $|\mu|<1$; this makes it possible to put $\mu=\cos\theta$ where $0<\theta<\pi$.

If the contour (L) is chosen so that it is symmetric with respect to the x-axis, and if the value of the part of the integral in formula (24) corresponding to the integration over the upper half of the contour is denoted by $A+iB$, the integral over the lower half is equal to $-(A-iB)=-A+iB$ (why?) and the whole integral is equal to $2iB$. Consequently, we can limit ourselves to integrating over the upper half and taking the imaginary part of the result. Let us represent the integral in form (22) where $f(z)=\frac{1}{z-\mu}$, $\nu=n$ and $\varphi(z)=\operatorname{Ln}\frac{z^2-1}{z-\mu}$ (here we take a concrete single-valued branch of the function $\varphi(z)$ which is multiple-valued and has the branch points $z=1$, $z=-1$ and $z=\mu$). We have

$$\varphi'(z)=\frac{z^2-2\mu z+1}{(z^2-1)(z-\mu)}$$

and hence the derivative vanishes at the points $z=\mu\pm\sqrt{\mu^2-1}=\cos\theta\pm\sqrt{\cos^2\theta-1}=e^{\pm i\theta}$ (these are the saddle-points of $|e^{\varphi(z)}|$). It is readily shown that the curve $\operatorname{Re}\varphi(z)=\operatorname{Re}\varphi(e^{i\theta})$ consists of two circles with centres at the points $z=\pm 1$ passing through the point $z_0=e^{i\theta}$ (Fig. 57). Since we have $\varphi''(z_0)=(ie^{i\theta}\sin\theta)^{-1}$ (check it up!), formula (24) yields the asymptotic representation of P_n for $n\to\infty$:

$$P_n(\cos\theta)\sim\frac{1}{2^{n+1}\pi i}\,2i\,\operatorname{Im}\left[\frac{1}{e^{i\theta}-\cos\theta}\sqrt{-\frac{2\pi}{n}ie^{i\theta}\sin\theta}\left(\frac{e^{2i\theta}-1}{e^{i\theta}-\cos\theta}\right)^n\right]$$

Since the radical in this formula must have a negative real part

(why?), it is equal to $\sqrt{\frac{2\pi}{n}\sin\theta}\,e^{i\left(\frac{3\pi}{4}+\frac{\theta}{2}\right)}$. Simplifying this expression we finally obtain, for $n\to\infty$, the formula

$$P_n(\cos\theta)\sim\frac{1}{2^n\pi}\operatorname{Im}\left[\frac{1}{i\sin\theta}\sqrt{\frac{2\pi}{n}\sin\theta}\,e^{i\left(\frac{3}{4}\pi+\frac{\theta}{2}\right)}(2e^{i\theta})^n\right]=$$
$$=\sqrt{\frac{2}{\pi n\sin\theta}}\sin\left(n\theta+\frac{\theta}{2}+\frac{\pi}{4}\right)$$

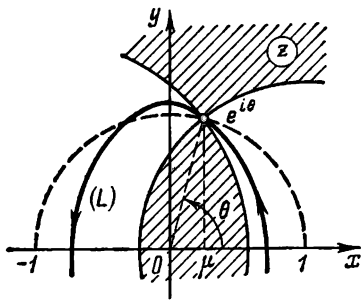


Fig. 57

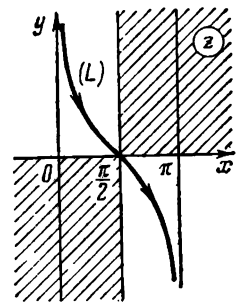


Fig. 58

Our considerations indicate that this asymptotic relation should be understood as

$$P_n(\cos\theta)=\sqrt{\frac{2}{\pi n\sin\theta}}\sin\left(n\theta+\frac{\theta}{2}+\frac{\pi}{4}\right)+O\left(n^{-\frac{3}{2}}\right)$$

but not as $\sqrt{\frac{2}{\pi n\sin\theta}}\sin\left(n\theta+\frac{\theta}{2}+\frac{\pi}{4}\right)\left(1+o\left(\frac{1}{n}\right)\right)$ (let the reader examine the difference between these expressions). This stipulation also refers to other cases when the dominant term of an asymptotic expansion has zeros.

Consider another example. Let us take the integral

$$H_\nu^{(1)}(x)=\frac{1}{\pi}\int_{(L)}e^{ix\sin z-i\nu z}\,dz \qquad (26)$$

in which $x>0$ and ν are real parameters and (L) is an arbitrary contour "passing from $\infty\,i$ to $\pi-\infty\,i$" (Fig. 58), that is having the straight lines $x=0$ and $x=\pi$ as asymptotes. It can readily be shown that integral (26) is convergent (and the rate of convergence is rather high) and independent of the particular choice of the path (L). The function $H_\nu^{(1)}(x)$ is called the **Hankel function of the first kind** (or the **Bessel function of the third kind**). It can be proved that it is equal to $J_\nu(x)+iY_\nu(x)$ where J_ν and Y_ν are, respectively, the **Bessel functions of the first and second**

kinds (e. g. see the definition of the Bessel functions in [107], Sec. XV.26).

We shall investigate the behaviour of function (26) for constant values of ν when $x \to \infty$. To this end, we replace ν by x in formula (26) and put $f(z) = e^{-ixz}$ and $\varphi(z) = i \sin z$. Then $z = \frac{\pi}{2}$ is a saddle-point, and the application of formula (24) (let the reader carry out the calculations) results in

$$H_\nu^{(1)}(x) = \sqrt{\frac{2}{\pi x}} \left\{ \exp\left[i \left(x - \frac{\nu\pi}{2} - \frac{\pi}{4} \right) \right] + O\left(\frac{1}{x} \right) \right\} \quad (x \to \infty) \quad (27)$$

In particular, from (27) we derive the asymptotic formulas

$$\left.\begin{aligned} J_\nu(x) &= \sqrt{\frac{2}{\pi x}} \cos\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4} \right) + O\left(x^{-\frac{3}{2}} \right) \\ Y_\nu(x) &= \sqrt{\frac{2}{\pi x}} \sin\left(x - \frac{\nu\pi}{2} - \frac{\pi}{4} \right) + O\left(x^{-\frac{3}{2}} \right) \end{aligned}\right\} (x \to \infty) \quad (28)$$

for the Bessel functions (e. g. see approximate graphs of the functions in [107], Fig. 299).

CHAPTER III

Operational Calculus

O. Heaviside (1850-1925), an English physicist, inaugurated in his *Electromagnetic Theory*, London, 1899, an *operational calculus*. He applied it to ordinary differential equations connected with electrotechnical problems but did not justify mathematically the meaning of operators occurring in the calculus. The interpretation of such operators based on the theory of analytic functions and connected with the theory of Fourier's integral transformation (e. g. see [107], § XVII.5) was given later by his successors. The operational calculus is widely used in modern applied mathematics, physics and engineering. There are a number of special monographs devoted to it (e. g. [16], [28], [30], [65], [72], [90]). We also refer the reader to the corresponding chapters of general courses [5], [81], [126].

§ 1. GENERAL THEORY

1. Laplace Transformation. The operational calculus is based on the so-called *Laplace transformation*. Let there be given a function $f(t)$ (which may be complex in the general case) of a real variable t, $0 \leqslant t < \infty$. Its **Laplace transform** is defined by the formula

$$F(p) = \int_0^\infty e^{-pt} f(t)\, dt \tag{1}$$

and is thus a function of the complex variable p. Formula (1) specifies the **Laplace transformation,** i. e. an operator (e. g. see [107], Sec. XIV.26) for which the function $f(t)$ is the **original (initial function, preimage, inverse image** or **inverse transform** of $F(p)$) while the function $F(p)$ is the **image (transform)** of $f(t)$. The relationship between an original function and its Laplace transform will be designated as $f(t) \longrightarrow F(p)$; the notation $f(t) \doteq F(p)$ and some others are also used.

Integral (1) being improper (e. g. see [107], § XIV.4), we must impose some conditions for its convergence. We shall suppose that *the function $f(t)$ is finite for all finite values of t* (although, as can be proved, it is sufficient to require that the function should be absolutely integrable) and that for $t \to \infty$ *it is bounded* or *has an exponential rate of growth*. The latter condition can be written in the general form

$$|f(t)| \leqslant Me^{s_0 t} \tag{2}$$

where M and s_0 are some constants. There are of course functions which increase, for $t \to \infty$, faster than any exponential expression (e. g. such is the function e^{t^2}) but they are not encountered in practical applications, and thus condition (2) is not too strong.

Putting $p = s + ir$ we can easily prove that integral (1) is absolutely convergent for $s > s_0$ if condition (2) is fulfilled. Indeed, we have

$$|e^{-pt} f(t)| = |e^{-st} e^{-irt} f(t)| = e^{-st} |f(t)| \leqslant Me^{-(s-s_0)t} \tag{3}$$

and the integral of the rightmost expression in (3) converges since $-(s - s_0) < 0$. Thus, the function $F(p)$ is defined in a region of the complex p-plane which contains the half-plane $\operatorname{Re} p > s_0$. This reveals an important distinction between Laplace's transformation and Fourier's transformation: for the latter it is required that the initial function vanish at infinity while for the former the original, due to the factor e^{-st}, can be of an exponential rate of growth at infinity.

Generally speaking, the admissible values of s_0 are different for different originals. Therefore, when we simultaneously deal with several initial functions we suppose that $\operatorname{Re} p$ is sufficiently large although we do not stipulate its precise value unless we are particularly interested in it.

It can readily be shown that for $\operatorname{Re} p > s_0$ integral (1), dependent on p as parameter, possesses a derivative with respect to p that can be computed according to the well-known *Leibniz rule* (e.g. see [107], § XIV.5) which holds for complex values of the parameter as well.

Indeed, if p varies in a small neighbourhood of a fixed point of the p-plane we can write down inequality of type (3) for all these values of p replacing s by the least value of $\operatorname{Re} p$ in this neighbourhood. Therefore, integral (1) and also the integral obtained by the formal differentiation of (1) with respect to the parameter p satisfy the conditions under which the differentiation is legitimate (e.g. see [107], Sec. XIV.21).

Thus, *$F(p)$ is a one-valued analytic function in the half-plane* $\operatorname{Re} p > s_0$.

Formula (1) indicates that the addition of preimages results in the addition of the images and that if an original is multiplied by a constant its transform is multiplied by the same constant. In other words, *the Laplace transformation determines a linear operator*. This implies (e.g. see [107], Sec. XVII.33) that if an original function depends on a parameter the derivative of its transform (which is also a function of this parameter) with respect to the parameter is the image of the derivative of the original with respect to the parameter.

2. Transforms of Simple Functions.

1. The transform of an exponential function is readily found:

$$e^{at} \longrightarrow \int_0^\infty e^{-pt} e^{at}\, dt = \frac{e^{(-p+a)t}}{-p+a}\Bigg|_{t=0}^{\infty} = 0 - \frac{1}{-p+a} = \frac{1}{p-a} \qquad (4)$$

Here the substitution of ∞ yields zero because, as was shown in Sec. 1, the value of $\mathrm{Re}\, p$ can be regarded as being arbitrarily large, and $e^{-\infty} = 0$.

In particular, putting $a = 0$ we obtain

$$1 \longrightarrow \frac{1}{p}$$

2. The left-hand and right-hand members of (4) contain a as a parameter and hence, differentiating with respect to the parameter (see Sec. 1), we receive

$$te^{at} \longrightarrow \frac{1}{(p-a)^2}, \quad t^2 e^{at} \longrightarrow \frac{1 \cdot 2}{(p-a)^3}, \quad \ldots, \quad t^n e^{at} \longrightarrow \frac{n!}{(p-a)^{n+1}}$$

In particular, we have

$$t^n \longrightarrow \frac{n!}{p^{n+1}} \qquad (5)$$

Formula (5) is readily generalized to any real value of n greater than -1. The condition $n > -1$ is necessary for the preimage (which has a singularity at $t = 0$ when $n < 0$) to be integrable. Thus, we have

$$t^n \longrightarrow \int_0^\infty e^{-pt} t^n\, dt = \int_0^\infty e^{-\tau} \frac{\tau^n}{p^n} \frac{d\tau}{p} = \frac{\Gamma(n+1)}{p^{(n+1)}} \qquad (6)$$

(when calculating (6) we have put $pt = \tau$ and used a well-known formula for the gamma function; e.g. see [107], Sec. XIV.17). In the change of variable $pt = \tau$ we can take real and positive values of p because the transform of t^n is an analytic function of p, and hence if it coincides with the right-hand side of (6) for real values of p it must be equal to it for imaginary values of p as well (why?).

3. If we consider the parameter a in formula (4) to be complex and put $a=\alpha+i\beta$ we receive

$$e^{\alpha t}\cos\beta t+ie^{\alpha t}\sin\beta t\longrightarrow\frac{1}{p-\alpha-i\beta}=\frac{p-\alpha}{(p-\alpha)^2+\beta^2}+i\frac{\beta}{(p-\alpha)^2+\beta^2}$$

It readily follows that the transform of the first summand on the left-hand side (denote this transform by $F_1(p)$) is the first summand on the right-hand side (denote it by $\Phi_1(p)$). The same relationship exists between the second summands ($F_2(p)$ and $\Phi_2(p)$, respectively).

In fact, the linearity of the Laplace transformation shows that the transform of the left-hand side is equal to $F_1(p)+iF_2(p)$, and thus $F_1(p)+iF_2(p)\equiv\Phi_1(p)+i\Phi_2(p)$. But all the functions involved assume real values for real p (why?), and therefore for real values of p we can write $F_1(p)=\Phi_1(p)$ and $F_2(p)=\Phi_2(p)$. But analytic functions coinciding for real p must coincide identically (see Sec. II.2.6), which implies what we set out to prove.

Consequently, we have

$$\begin{aligned}e^{\alpha t}\cos\beta t&\longrightarrow\frac{p-\alpha}{(p-\alpha)^2+\beta^2}\\ e^{\alpha t}\sin\beta t&\longrightarrow\frac{\beta}{(p-\alpha)^2+\beta^2}\end{aligned}\qquad(7)$$

4. Let $f(t)$ be a periodic function with period $T>0$. Then its Laplace transform can be computed according to the formula (check it up!)

$$f(t)\longrightarrow\sum_{n=0}^{\infty}\int_{nT}^{(n+1)T}e^{-pT}f(t)\,dt=|\,nT+\tau=t\,|=$$

$$=\sum_{n=0}^{\infty}\int_{0}^{T}e^{-pnt}e^{-p\tau}f(\tau+nT)\,d\tau=$$

$$=\int_{0}^{T}e^{-p\tau}f(\tau)\,d\tau\sum_{n=0}^{\infty}(e^{-pT})^n=\frac{1}{1-e^{-pT}}\int_{0}^{T}e^{-pt}f(t)\,dt$$

5. Let us consider the *Laplace transform of the delta function* $\delta(t)$ (e.g. see [107], Sec. XIV.25). The singularity of $\delta(t)$ being concentrated at $t=0$, we must stipulate what "portion" of this singularity is integrated in (1). Let us suppose that in (1) the integration extends from -0 to ∞ (the same result is obtained if we formally replace $\delta(t)$ by $\delta(t-0)$ and take the lower limit 0). Then we get

$$\delta(t)\longrightarrow\int_{-0}^{+0}e^{-pt}\delta(t)dt+\int_{+0}^{\infty}e^{-pt}\delta(t)\,dt=1\cdot\int_{-0}^{+0}\delta(t)\,dt+0=1\qquad(8)$$

Now we can compile the following table of the simplest transforms which most often occur in applications:

$f(t)$	$F(p)$
1	$\frac{1}{p}$
t^n	$\frac{n!}{p^{n+1}}=\frac{\Gamma(n+1)}{p^{n+1}}\ (n>-1)$
e^{at}	$\frac{1}{p-a}$
$t^n e^{at}$	$\frac{n!}{(p-a)^{n+1}}=\frac{\Gamma(n+1)}{(p-a)^{n+1}}\ (n>-1)$
$e^{\alpha t}\cos\beta t$	$\frac{p-\alpha}{(p-\alpha)^2+\beta^2}$
$e^{\alpha t}\sin\beta t$	$\frac{\beta}{(p-\alpha)^2+\beta^2}$
$\delta(t)$	1

For more extensive tables we refer the reader to [27] and [29].

3. Basic Properties of the Laplace Transformation.

1. Let $f(t)\longrightarrow F(p)$. Then for constant $k>0$ we have

$$f(kt)\longrightarrow\int_0^\infty e^{-pt}f(kt)\,dt=\int_0^\infty e^{-p\frac{\tau}{k}}f(\tau)\frac{d\tau}{k}=\frac{1}{k}F\left(\frac{p}{k}\right)$$

(here we have made the substitution $kt=\tau$; let the reader check up the calculations).

2. Let $f(t)\longrightarrow F(p)$. Then, for constant values of τ, we have

$$f(t-\tau)\longrightarrow\int_0^\infty e^{-pt}f(t-\tau)\,dt=\int_{-\tau}^\infty e^{-p(s+\tau)}f(s)\,ds\qquad(t-\tau=s)$$

Now putting $s=t$ we finally obtain

$$f(t-\tau)=\int_{-\tau}^0 e^{-pt}e^{-p\tau}f(t)\,dt+\int_0^\infty e^{-pt}e^{-p\tau}f(t)\,dt=$$

$$=e^{-p\tau}F(p)+e^{-p\tau}\int_{-\tau}^0 e^{-pt}f(t)\,dt$$

This result becomes particularly simple when $\tau>0$ and $f(t)\equiv 0$ for $t<0$. In this case we receive

$$f(t-\tau)\longrightarrow e^{-p\tau}F(p)$$

The latter formula expresses the so-called **delay theorem**, the name being accounted for by the fact that the independent variable t is most often interpreted as time, and the function $f(t-\tau)$ then expresses the same law of variation as $f(t)$ but with the initial moment of time $t=\tau$ instead of $t=0$.

3. Let $f(t) \longrightarrow F(p)$. If c is an arbitrary complex constant we have

$$e^{ct} f(t) \longrightarrow \int_0^\infty e^{-pt} e^{ct} f(t)\, dt = \int_0^\infty e^{-(p-c)t} f(t)\, dt = F(p-c)$$

This is the so-called **shift theorem.**

4. Let $f(t) \longrightarrow F(p)$. Integrating by parts we write

$$f'(t) \longrightarrow \int_0^\infty e^{-pt} f'(t)\, dt = \int_0^\infty e^{-pt}\, df(t) =$$

$$= e^{-pt} f(t) \Big|_{t=0}^{\infty} - \int_0^\infty f(t)\, e^{-pt}(-p)\, dt = pF(p) - f(+0) \qquad (9)$$

When substituting the upper limit we obtain zero because, according to the hypothesis, even if the function $f(t)$ increases for $t \longrightarrow \infty$ it is of an exponential rate of growth, and, as was indicated in Sec. 1, the real part of p can be regarded as being arbitrarily large. On the right-hand side we have written $f(+0)$ because the function $f(t)$ may have a jump discontinuity at the point $t=0$.

Differentiating repeatedly we obtain

$$f''(t) = [f'(t)]' \longrightarrow p[pF(p) - f(+0)] - f'(+0) =$$
$$= p^2 F(p) - pf(+0) - f'(+0)$$

and, generally,

$$f^{(n)}(t) \longrightarrow p^n F(p) - p^{n-1} f(+0) - p^{n-2} f'(+0) - \ldots - f^{(n-1)}(+0)$$

For the case $f(+0)=0$ formula (9) takes a particularly simple form: $f'(t) \longrightarrow pF(p)$; the same refers to the formulas for the subsequent derivatives.

Formula (9) can be given the following interpretation. Let us consider the function $f(t)$ subjected to the Laplace transformation as being defined over the whole t-axis (i.e. not only for $t \geqslant 0$) and satisfying the condition $f(t) \equiv 0$ for $t < 0$. Besides, let the lower limit of integration in formula (1) be $-\infty$ or, equivalently, -0. Then formula (9) can always be written as $f'(t) \longrightarrow pF(p)$. But if the value $f(+0)$ is nonzero and finite the derivative $f'(t)$ includes the term $f(+0)\,\delta(t)$ containing the delta function (e.g. see [107], Sec. XIV.25). Let us denote by $[f'(t)]$ the function

$f'(t)$ with that term discarded. Then we have

$$pF(p) \leftarrow f'(t) = [f'(t)] + f(+0)\,\delta(t)$$

Returning to Example 5 in Sec. 2 (in which we now do not stipulate what "portion" of the delta function is integrated) we see that the transform of the function $[f'(t)]$ is equal to the expression $pF(p) - f(+0)$. If we similarly consider $f''(t)$, the expression $\delta'(t)$ appears (e.g. see [107], Sec. XIV.27), etc. Using functions of this kind we can simplify many expressions involved although then some further facts of the theory below must be given another interpretation. In what follows we shall not resort to this approach.

Let the reader prove that if $f(t) \rightarrow F(p)$ then

$$\int_0^t f(\tau)\,d\tau \rightarrow \frac{1}{p} F(p) \tag{10}$$

5. Differentiating both sides of formula (1) with respect to p we receive (check it up!)

$$tf(t) \rightarrow -F'(p)$$

It follows that if the function $\frac{f(t)}{t}$ is bounded or at least absolutely integrable in the vicinity of the point $t=0$, then, denoting its Laplace transform by $\Phi(p)$, we can write $f(t) = t\,\frac{f(t)}{t} \rightarrow$ $\rightarrow -\Phi'(p)$, that is

$$\Phi'(p) = -F(p) \text{ and therefore } \Phi(p) = -\int F(p)\,dp$$

But formula (1) shows that for real $p = +\infty$ the transform of any function vanishes, and therefore we finally obtain

$$\frac{f(t)}{t} \rightarrow -\int_{+\infty}^{p} F(q)\,dq \tag{11}$$

6. The **convolution** of two functions $f_1(t)$ and $f_2(t)$ defined for $0 \leqslant t < \infty$ is a new function of t designated as $f_1(t) * f_2(t)$ and determined by the formula

$$f_1(t) * f_2(t) = \int_0^t f_1(\tau)\, f_2(t-\tau)\,d\tau \quad (0 \leqslant t < \infty) \tag{12}$$

It should be noted that the value of function (12) for any $t > 0$ is specified by the values of the functions f_1 and f_2 on the whole interval $0 \leqslant \tau \leqslant t$, and therefore it would be more precise to denote the convolution as $(f_1 * f_2)(t)$ instead of $f_1(t) * f_2(t)$ but the latter notation is more convenient for practical purposes.

Making the change of variable $\tau = t - \tau_1$ in the right-hand side of (12) we obtain (check it up!)

$$f_1(t) * f_2(t) = f_2(t) * f_1(t) \tag{13}$$

It can also be shown directly that the other ordinary properties of multiplication also hold for convolutions, but this will become clear on the basis of some other considerations given below.

As an example, we compute the convolution of t^2 and t:

$$t^2 * t = \int_0^t \tau^2 (t-\tau)\, d\tau = \left(t\frac{\tau^3}{3} - \frac{\tau^4}{4} \right)\Big|_{\tau=0}^{t} = \frac{1}{12} t^4$$

Let now $f_1(t) \longrightarrow F_1(p)$ and $f_2(t) \longrightarrow F_2(p)$. Then we have

$$f_1(t) * f_2(t) \longrightarrow \int_0^\infty e^{-pt} \left(\int_0^t f_1(\tau) f_2(t-\tau)\, d\tau \right) dt =$$

$$= \int_0^\infty dt \int_0^t e^{-pt} f_1(\tau) f_2(t-\tau)\, d\tau$$

Reversing the order of integration (e.g. see [107], Sec. XVI.9) we obtain

$$f_1(t) * f_2(t) \longrightarrow \int_0^\infty d\tau \int_\tau^\infty e^{-pt} f_1(\tau) f_2(t-\tau)\, dt =$$

$$= |\, t - \tau = t_1 \,| = \int_0^\infty f_1(\tau)\, d\tau \int_0^\infty e^{-p\tau} e^{-pt_1} f_2(t_1)\, dt_1 =$$

$$= \int_0^\infty e^{-p\tau} f_1(\tau)\, d\tau \int_0^\infty e^{-pt_1} f_2(t_1)\, dt_1 = F_1(p)\, F_2(p)$$

Thus, *the Laplace transform of the convolution of two functions is equal to the product of the transforms of these functions.* This implies that the properties of multiplication are extended to convolutions; for instance, if we pass to the transforms in formula (13) we see that this formula is directly implied by the commutativity of multiplication. We also see from formula (8) that the role of the delta functions for convolutions is similar to that of unity for multiplication, i.e. the convolution of a function and the delta function coincides with the former function which thus remains invariable; by the way, this can also be easily proved directly.

Using formulas of Sec. 2 and applying the above properties we can derive many useful formulas for Laplace transforms but here we shall not dwell on these questions.

4. Laplace's Inverse Transformation. As is known (e.g. see [107] formulas (XVII.138) and (XVII.141)), if $\varphi(t)$ is a given function, it is connected with its **Fourier transform** $\hat{\varphi}(t)$ by the formulas

$$\hat{\varphi}(r)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\varphi(t)e^{-irt}dt,\quad \varphi(t)=\int_{-\infty}^{\infty}\hat{\varphi}(r)e^{irt}dr \tag{14}$$

which hold for any absolutely integrable function $\varphi(t)$. To apply these formulas to transformation (1) we put $p=s+ir$ and

$$\varphi(t)=\begin{cases}2\pi f(t)e^{-st} & \text{for } t>0\\ 0 & \text{for } t<0\end{cases} \tag{15}$$

where s is fixed and sufficiently large (we must have $s>s_0$ where s_0 enters into inequality (2)). Then we can write

$$\hat{\varphi}(r)=\frac{1}{2\pi}\int_{-\infty}^{0}\varphi(t)e^{-irt}dt+\frac{1}{2\pi}\int_{0}^{\infty}\varphi(t)e^{-irt}dt=$$

$$=\frac{1}{2\pi}\int_{0}^{\infty}2\pi f(t)e^{-st}e^{-irt}dt=\int_{0}^{\infty}e^{-(s+ir)t}f(t)dt=F(s+ir) \tag{16}$$

(we see that notation (15) has been used here for writing integral (1) in the form of the first integral (14)). The second formula (14) now implies that

$$\int_{-\infty}^{\infty}F(s+ir)e^{irt}dt=\begin{cases}2\pi f(t)e^{-st} & \text{for } t>0\\ 0 & \text{for } t<0\end{cases} \tag{17}$$

Formula (17) shows that for $t>0$ we have

$$f(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{(s+ir)}F(s+ir)dr \tag{18}$$

If r varies from $-\infty$ to ∞ the point $p=s+ir$ describes a straight line in the complex p-plane which is parallel to the imaginary axis. In this case we shall (conditionally) say that p varies from $s-i\infty$ to $s+i\infty$. Here we have $dp=i\,dr$, and therefore it follows from (18) that

$$f(t)=\frac{1}{2\pi i}\int_{s-i\infty}^{s+i\infty}e^{pt}F(p)dp\qquad(t>0) \tag{19}$$

This is the **inversion formula** for the Laplace transformation.

It should be noted that s on the right-hand side of (19) is an arbitrary sufficiently large constant number. The above proof shows that the integral is independent of the choice of s (let the reader

give an independent proof of this fact by establishing, on the basis of formula (2), the inequality

$$|F(s+ir)| \leqslant \frac{M}{s-s_0} \qquad (s > s_0)$$

and then applying Cauchy's theorem from Sec. II.2.2).

Formula (17) also indicates that for $t<0$ the right-hand side of (19) is equal to zero. Taking into account the analogous property of Fourier's transformation (e.g. see [107], the end of Sec. XVII.32) we thus conclude that for $t=0$ the right-hand side of (19) is equal to $\frac{1}{2}f(+0)$. It also follows that not only every original uniquely determines its Laplace transform but also any given transform specifies a unique initial function, that is the operator (mapping) defined by formula (1) is one-to-one.

It should be noted that integral (19) may turn out to be divergent because of the infinite limits of integration $p=s\pm i\infty$. It can be proved that in this case it is possible to interpret it in the sense of Cauchy's principal value (e.g. see [107], Sec. XIV.19).

As is known (e.g. see [107], the end of Sec. XVII.33), for the Fourier transformation we have the *Parseval relation*. For the functions φ and $\hat{\varphi}$ considered above Parseval's relation takes the form (check it up!)

$$2\pi \int_0^\infty |f(t)|^2 e^{-2st}\,dt = \int_{-\infty}^{\infty} |F(s+ir)|^2\,dr \qquad (s \geqslant s_0)$$

which is convenient for some applications.

In some problems we are given a function $F(p)$ for which the corresponding function $f(t)$ must be found. Then there arises the question of what functions $F(p)$ can serve as Laplace transforms. As was shown, the function $F(p)$ must be analytic in every half-plane $\operatorname{Re} p \geqslant s$ $(s > s_0)$. Besides, it can readily be proved, on the basis of (1) and (2), that if $f(t)$ is an ordinary function (i.e. not a generalized function of the type of $\delta(t)$) we have the relation $\lim_{|p|\to\infty} F(p)=0$ in this half-plane. It turns out that, generally speaking, these conditions are sufficient for the function $F(p)$ to serve as a Laplace transform. Indeed, let us additionally suppose that $\int_{-\infty}^{\infty} |F(s+ir)|\,dr < \infty$ (this condition can in fact be weakened) and define the function $f(t)$ for $t>0$ by means of formula (19). But if $t<0$, integral (19) is equal to zero (which can easily be proved if we put $t=-\omega$, $p=s-z$, pass, as it was done in Sec. II.3.4, from the integration along the straight line to the integration over the left semicircle in the z-plane and apply

Jordan's lemma, i.e. formula (II.3.22)). Hence, we can pass to formulas (17) and then, by means of (14) and (15), to equality (16) which indicates that $F(p)$ is the Laplace transform of $f(t)$.

Consider an example. Take the function

$$F(p)=\frac{1}{p}e^{-\sqrt{p}} \tag{20}$$

where we mean the branch of the square root satisfying the normalization condition $\sqrt{p}\,|_{p=1}=1$. From (19) we receive the inverse transform

$$f(t)=\frac{1}{2\pi i}\int_{s-i\infty}^{s+i\infty}e^{pt}\frac{1}{p}e^{-\sqrt{p}}\,dp \qquad (t>0) \tag{21}$$

To compute the integral in (21) (we take it without the factor $\frac{1}{2\pi i}$) we apply the theory of integration of analytic functions (Sec. II.3.4). If the p-plane is cut along the negative real half-axis the integrand can be regarded as single-valued (why?). Let us first take the integral over the contour shown in Fig. 59 where the curvilinear parts are circular arcs of radii R and ρ ($R\to\infty$ and $\rho\to 0$). Then, by (21), the integral along portion I of the contour tends to $2\pi i\,f(t)$. The integrals along portions II and VI tend to zero; this is implied by Jordan's lemma in which we must put $\omega=t$, $z=p-s$ and take into account that the semicircular arc (M_R) (formula (II.3.22)) can be replaced by its portion. The integral along portion IV tends to $-2\pi i$ (prove this assertion by expanding the exponential functions under the integral sign into series). On the arcs III and V we have $p=-\xi$, $0<\xi<\infty$ and therefore, respectively, $\sqrt{p}=i\sqrt{\xi}$ and $\sqrt{p}=-i\sqrt{\xi}$. Consequently, the sum of the corresponding integrals tends to

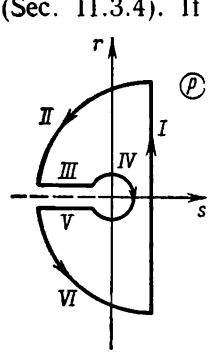


Fig. 59

$$\int_0^\infty e^{-t\xi}\frac{1}{-\xi}\left(e^{i\sqrt{\xi}}-e^{-i\sqrt{\xi}}\right)(-d\xi)=2i\int_0^\infty e^{-t\xi}\frac{\sin\sqrt{\xi}}{\xi}\,d\xi$$

But the integral over the whole closed contour is equal to zero, and hence, passing to the limit and cancelling by $2i$, we obtain

$$\pi f(t)-\pi+\int_0^\infty e^{-t\xi}\frac{\sin\sqrt{\xi}}{\xi}\,d\xi=0 \tag{22}$$

In order to compute the integral on the left-hand side of (22) (denote it by $I(t)$) we differentiate it with respect to the parameter t:

$$\frac{dI}{dt} = -\int_0^\infty e^{-t\xi} \sin\sqrt{\xi}\, d\xi = \left| t\xi = x^2 \right| =$$

$$= -\frac{1}{2}\int_{-\infty}^{\infty} e^{-x^2} \sin\frac{x}{\sqrt{t}} \cdot 2\frac{x}{t}\, dx = -\frac{1}{t}\operatorname{Im}\int_{-\infty}^{\infty} e^{-x^2+\frac{ix}{\sqrt{t}}}\, x\, dx =$$

$$= -\frac{1}{t}\operatorname{Im}\int_{-\infty}^{\infty} e^{-\left(x-\frac{i}{2\sqrt{t}}\right)^2-\frac{1}{4t}}\left[\left(x-\frac{i}{2\sqrt{t}}\right)+\frac{i}{2\sqrt{t}}\right] dx$$

If we open the square brackets and split the last expression into two integrals we find that the first integral can be computed by the Newton-Leibniz formula (the corresponding indefinite integral is expressed in terms of elementary functions) and is equal to zero while the second integral is equal to

$$\frac{i}{2\sqrt{t}} e^{-\frac{1}{4t}} \int_{-\infty}^{\infty} e^{-x^2} dx = \frac{i}{2\sqrt{t}} e^{-\frac{1}{4t}} \sqrt{\pi} \tag{23}$$

(the passage from the integral with exponent $-\left(x-\frac{i}{2\sqrt{t}}\right)^2$ to integral (23) can be performed as in [107], page 749, or by applying Cauchy's theorem from Sec. II.2.2). Since $I(\infty) = 0$, we have

$$I(t) = -\int_\infty^t \frac{\sqrt{\pi}}{2} \tau^{-\frac{3}{2}} e^{-\frac{1}{4\tau}} d\tau.$$

Therefore, from (22) we obtain

$$f(t) = 1 - \frac{1}{2\sqrt{\pi}} \int_t^\infty \tau^{-\frac{3}{2}} e^{-\frac{1}{4\tau}} d\tau = \left| \frac{1}{4\tau} = s^2 \right| =$$

$$= 1 - \frac{2}{\sqrt{\pi}} \int_0^{\frac{1}{2}\sqrt{t}} e^{-s^2} ds = 1 - \frac{2}{\sqrt{\pi}} \operatorname{Erf}\left(\frac{1}{2\sqrt{t}}\right)$$

(the symbol "Erf" means "*error function*"; e.g. see [107], Sec. XIV.12). The latter expression is the inverse transform of function (20). Using Property 4, Sec. 3, we also deduce

$$\frac{1}{2\sqrt{\pi}} t^{-\frac{3}{2}} e^{-\frac{1}{4t}} \longrightarrow e^{-\sqrt{p}} \tag{24}$$

Exercise. Prove that

$$\frac{1}{\sqrt{\pi t}} e^{-\frac{1}{4t}} \rightarrow \frac{1}{\sqrt{p}} e^{-\sqrt{p}} \tag{25}$$

5. Expanding the Original into a Series. As was shown in Sec. 1, the function $F(p)$ is one-valued and analytic in the half-plane $\operatorname{Re} p > s_0$. In many cases this function turns out to be one-valued analytic in the whole p-plane except at a number of isolated singular points. But, in the general case the function $F(p)$ does not necessarily possess this property because, when continued to the entire p-plane, it may have branch points (Sec. II.1.10) as in Example (20). But here we shall suppose that the *function $F(p)$ is single-valued in the whole p-plane.*

Besides, let us suppose that $F(p)$ has a finite number of singular points $p_1, p_2, \ldots, p_n$ and tends to zero when p approaches infinity in an arbitrary way (in the preceding sections we only showed that $F(p) \underset{p \to \infty}{\longrightarrow} 0$ for $\operatorname{Re} p \geqslant s$). Then integral (19) can be computed by means of the residue theorem and Jordan's lemma (see formula (II.3.22)) as described in Sec. II.3.4. This results in

$$f(t) = \frac{1}{2\pi i} 2\pi i \sum_{k=1}^{n} \operatorname*{Res}_{p=p_k} [e^{pt} F(p)] = \sum_{k=1}^{n} \operatorname*{Res}_{p=p_k} [e^{pt} F(p)] \tag{26}$$

(check it up!).

In particular, if $F(p)$ is a proper rational fraction the above requirements are obviously fulfilled (why?). If

$$F(p) = \frac{P(p)}{Q(p)}$$

where all the roots $p_1, p_2, \ldots, p_n$ of the polynomial $Q(p)$ are simple and its degree n exceeds the degree of the polynomial $P(p)$ formula (26) becomes particularly simple. Indeed, in this case we can apply formula (II.3.5) for computing the residues and thus find

$$f(t) = \sum_{k=1}^{n} \operatorname*{Res}_{p=p_k} \frac{P(p) e^{pt}}{Q(p)} = \sum_{k=1}^{n} \frac{P(p_k)}{Q'(p_k)} e^{p_k t}$$

If the function $F(p)$ has an infinitude of isolated singular points $p_1, p_2, \ldots$ the Bolzano-Weierstrass theorem (Sec. II.2.6) obviously implies that $p_k \underset{k \to \infty}{\longrightarrow} \infty$. Then of course the condition $F(p) \underset{p \to \infty}{\longrightarrow} 0$ cannot be fulfilled (why?) but it may hold for a sequence of left semicircular arcs with centre at the point $p = s$ and infinitely increasing radii. Then, reasoning as in Sec. II.3.4, we can represent the function $f(t)$ as a sum of an infinite series similar to (26).

There is another important special case when the inverse transform of a function $F(p)$ can be expanded into a simple series, namely, when $F(p)$ is analytic at $p=\infty$ (see Sec. II.3.6). Indeed, at $p=\infty$ we then can write the expansion

$$F(p)=\sum_{k=0}^{\infty} c_{-k}p^{-k} \tag{27}$$

and, since $F(\infty)=0$ (why?), we have $c_0=0$, that is the summation is in fact carried out beginning with $k=1$. By formula (5), the inverse transform of the general term of series (27) is equal to $c_{-k}\frac{t^{k-1}}{(k-1)!}$, and therefore, since the inverse transform of a sum is equal to the sum of the inverse transforms of the summands, we obtain

$$f(t)=\sum_{k=1}^{\infty} c_{-k}\frac{t^{k-1}}{(k-1)!} \tag{28}$$

(here we have used the linearity property in the case of an infinite number of summands, which of course requires justification; a more detailed investigation which we shall not present here shows that this is legitimate).

Consider an example. Let

$$F(p)=\frac{1}{\sqrt{p^2+1}} \tag{29}$$

where the square root is considered to be positive for real values of $p>0$. The analytic continuation of this function to the entire p-plane is a two-valued function; but if we limit ourselves to the neighbourhood of $p=\infty$, we find that the branch of the function $F(p)$ we are interested in is single-valued and can be expanded by means of the binomial series in powers of $\frac{1}{p^2}$:

$$F(p)=\frac{1}{p}(1+p^{-2})^{-\frac{1}{2}}=\frac{1}{p}\left(1+\frac{\left(-\frac{1}{2}\right)}{1}p^{-2}+\right.$$
$$\left.+\frac{\left(-\frac{1}{2}\right)\cdot\left(-\frac{3}{2}\right)}{1\cdot 2}p^{-4}+\frac{\left(-\frac{1}{2}\right)\cdot\left(-\frac{3}{2}\right)\cdot\left(-\frac{5}{2}\right)}{1\cdot 2\cdot 3}p^{-6}+\ldots\right)$$

Opening the brackets and taking advantage of formula (28) we find the inverse transform:

$$f(t)=1-\frac{1}{1!\,2^1}\cdot\frac{t^2}{1\cdot 2}+\frac{1\cdot 3}{2!\cdot 2^2}\cdot\frac{t^4}{1\cdot 2\cdot 3\cdot 4}-\frac{1\cdot 3\cdot 5}{3!\cdot 2^3}\cdot\frac{t^6}{1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6}+\ldots$$

Cancelling the fractions we arrive, after simple transformations, at the expression

$$f(t)=1-\frac{t^2}{(1!)^2\cdot 2^2}+\frac{t^4}{(2!)^2 2^4}-\frac{t^6}{(3!)^2\cdot 2^6}+\ldots$$

But this is nothing but the expansion of the Bessel function $J_0(t)$ (e.g. see [107], formula (XV.173)) which thus is the inverse transform of function (29).

It should be noted that a Laplace transform, i.e. a function of form (1), by far not always possesses an expansion into a convergent series of form (27) but, as a rule, it has an asymptotic expansion (Sec. II.4.1) of the same form valid for $\operatorname{Re} p \geqslant s > s_0$, $|p| \to \infty$. This can be proved as in Sec. II.4.3; for instance, if for small t the function $f(t)$ admits of a Taylor expansion

$$f(t) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} t^k \tag{30}$$

we can perform integration by parts in integral (1) infinitely many times and thus obtain (check it up!) the expansion

$$F(p) \sim \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{p^{k+1}} \tag{31}$$

Conversely, from this expansion we can easily restore series (30) (that is the values of $f(t)$ for small t) but not the values of $f(t)$ for all admissible values of t because any arbitrary variation of the values of the function for finite and large values of t does not affect expansion (3).

In particular, from (31) we conclude that

$$f(0) = \lim_{p \to \infty} [pF(p)] \qquad (\operatorname{Re} p \geqslant s)$$

For $s_0 \leqslant 0$ this formula, as is readily proved, can be inverted:

$$f(\infty) = \lim_{p \to 0} [pF(p)] \tag{32}$$

Indeed, we can write

$$pF(p) = \int_0^{\infty} f(t)\,(pe^{-pt}\,dt) =$$

$$= f(0) + \int_0^{\infty} e^{-pt} f'(t)\,dt \underset{p \to 0}{\longrightarrow} f(0) + \int_0^{\infty} f'(t)\,dt = f(\infty)$$

There is an interesting application of formula (32). Let us first apply property (10) to formula (11); this results in

$$\int_0^t \frac{f(\tau)}{\tau}\,d\tau \longrightarrow \frac{1}{p} \int_p^{\infty} F(q)\,dq$$

whence, by (32), we obtain

$$\int_0^{\infty} \frac{f(t)}{t}\,dt = \int_0^{\infty} F(p)\,dp$$

(we have replaced τ by t and q by p). This formula enables us to compute many other integrals for which the corresponding indefinite integrals are not reducible to elementary functions. For instance, from the second formula (7) we derive, for $\alpha = -\gamma \leqslant 0$, the relation

$$\int_0^{\infty} \frac{e^{-\gamma t}}{t} \sin \beta t\,dt = \int_0^{\infty} \frac{\beta}{(p+\gamma)^2+\beta^2}\,dp = \arctan \frac{p+\gamma}{\beta}\Big|_0^{\infty} = \frac{\pi}{2} - \arctan \frac{\gamma}{\beta}$$

There are methods which make it possible to investigate the behaviour of $f(t)$ for $t \longrightarrow \infty$ in more detail. In particular, we state (without proof) the following result. Let the function $F(p)$, when continued to the half-plane $\operatorname{Re} p \leqslant s_0$, have, among its singularities, only one singular point p_0 with the greatest real part (i.e. $F(p)$ is a single-valued analytic function in the half-plane $\operatorname{Re} p \geqslant \operatorname{Re} p_0$ except at the point p_0). Suppose that in the vicinity of the point p_0 the function $F(p)$ has the expansion

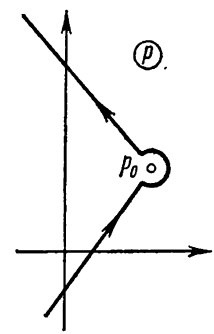


Fig. 60

$$F(p) = \sum_{k=0}^{\infty} c_k (p-p_0)^{\lambda_k}$$

where all the numbers λ_k are real and $\lambda_0 < \lambda_1 < \ldots \longrightarrow \infty$. Let us also suppose that in inversion formula (19) it is permissible to replace the path of integration by the one shown in Fig. 60. Then, for $t \longrightarrow \infty$, we have the asymptotic expansion

$$f(t) \sim e^{p_0 t} \sum_{k=0}^{\infty} \frac{c_k}{\Gamma(-\lambda_k)}\, t^{-\lambda_k - 1} \tag{33}$$

which is understood in the sense of Sec. II.4.1. If the function $F(p)$ has several singular points whose real parts coincide with the greatest, the corresponding expansions of form (33) must be added together.

6. Numerical Computation of the Original. Inversion formula (19) or the equivalent formula (18) can be used for numerical computation of the inverse transform $f(t)$ corresponding to a given transform $F(p)$ by applying numerical integration (e.g. see [107], Sec. XIV.13), which results in an approximate representation of

$f(t)$ in the form of a product of e^{st} by a sum of Fourier's series. As is known (e.g. see [107], Sec. XVII.25), the higher the degree of smoothness of the expanded function (i.e. the greater the number of continuous derivatives it possesses), the greater the speed of convergence of its Fourier series. The discontinuities of the function $f(t)$ and of its derivatives can be found by investigating the behaviour of the function $F(p)$; for instance, if $f(t_0+0)-f(t_0-0)=h\neq 0$ $(t_0>0)$ it can easily be shown that $F(p)$ includes a term of the form $h\frac{e^{-t_0p}}{p}$; the discontinuities of $f'(t)$ can be determined in the same manner, etc. Subtracting from $F(p)$ the terms of this kind we can obtain a function whose inverse transform is sufficiently smooth. Furthermore, as was shown, in the general case integral (18) may have a jump at the point $t=0$, and this can also result in the appearance of discontinuities on the t-axis. In this case we can improve the situation by subtracting from $F(p)$ certain terms whose inverse transforms are known; for instance, if $f(+0)$ is finite and different from zero the principal term in the expansion of $F(p)$ for $p\to\infty$ (for large values of $\operatorname{Re} p$) is of the form $\frac{c}{p}$ $(c=f(+0))$ in accordance with (6). Besides, if $f(+0)=0$ (i.e. $(pF(p))|_{p=\infty}=0$) and $f'(+0)$ is finite (i.e. $(p^2F(p))|_{p=\infty}$ is finite) we can continue integral (18) (taken without the factor e^{st}) in an odd fashion from the semiaxis $t\geqslant 0$ to the semiaxis $t\leqslant 0$, which means that we pass to the function

$$\tilde{f}(t)=\frac{1}{2\pi}e^{st}\left[\int_{-\infty}^{\infty}e^{irt}F(s+ir)\,dr-\int_{-\infty}^{\infty}e^{-irt}F(s+ir)\,dr\right]=$$

$$=\frac{i}{\pi}e^{st}\int_{-\infty}^{\infty}\sin rt\cdot F(s+ir)\,dr$$

Then the functions $\tilde{f}(t)$ and $\tilde{f}'(t)$ no longer have discontinuities at the point $t=0$.

Now let us suppose that such a continuation has been carried out and apply the trapezoid formula with the nodes $r=\frac{2\pi}{T}n$ where $T>0$ is a fixed number and $n=0,\ \pm1,\ \pm2,\ \ldots$. We then obtain an approximate expression of $\tilde{f}(t)$ (and therefore of $f(t)$) for $t>0$ in the form of a product of an exponential function by a sum of Fourier's series (check it up!):

$$f(t)\approx\frac{2i}{T}e^{st}\sum_{n=1}^{\infty}\left[F\left(s+i\frac{2\pi n}{T}\right)-F\left(s-i\frac{2\pi n}{T}\right)\right]\sin\frac{2\pi n}{T}t\qquad(t>0)$$

Of course, the chosen magnitude of T should be considerably greater than the values of t for which we compute $f(t)$.

Let us consider one more method of approximate computation of the original function which is based on manipulating the values of $F(p)$ for real p. We assume that $s_0 < 0$ (we can always reduce the problem to this case by applying the shift theorem) and that $f(+0)=0$ (see above). Suppose that there is a real number $q > 0$ for which the values

$$b_n = F((2n+1)q) \qquad (n=0,\ 1,\ \ldots)$$

are known. Let us make the change of variable $e^{-qt}=\cos\theta$ in integral (1) and put

$$f(t)=f\left(-\frac{1}{q}\ln\cos\theta\right)=\tilde{f}(\theta) \qquad \left(0\leqslant\theta\leqslant\frac{\pi}{2}\right)$$

$\left(\text{here we have } \tilde{f}(0)=\tilde{f}\left(\frac{\pi}{2}\right)=0\right)$. Then we obtain (check it up!)

$$qb_n=\int_0^{\frac{\pi}{2}}(\cos\theta)^{2n}\sin\theta\,\tilde{f}(\theta)\,d\theta \tag{34}$$

It is clear that the problems of constructing the functions $f(t)$ and $\tilde{f}(\theta)$ are equivalent.

Let us continue the function $\tilde{f}(\theta)$ in such a way that the extended function (which again is denoted as $\tilde{f}(\theta)$) is odd and its graph is symmetric with respect to the straight line $\theta=\frac{\pi}{2}$ $\left(\text{that is } \tilde{f}\left(\frac{\pi}{2}-\theta\right)\equiv\tilde{f}\left(\frac{\pi}{2}+\theta\right)\right)$. Then the function $\tilde{f}(\theta)$ is periodic with period 2π and can be expanded into Fourier's series

$$\tilde{f}(\theta)=\sum_{n=0}^{\infty}c_n\sin(2n+1)\theta \tag{35}$$

(why?). In particular, this formula holds for $0\leqslant\theta\leqslant\frac{\pi}{2}$. Substituting (35) into (34) we obtain, after simple calculations, the triangular system of linear equations

$$\left.\begin{aligned}
&c_0=\frac{4}{\pi}qb_0\\
&c_0+c_1=\frac{4^2}{\pi}qb_1\\
&2c_0+3c_1+c_2=\frac{4^3}{\pi}qb_2\\
&5c_0+9c_1+5c_2+c_3=\frac{4^4}{\pi}qb_3\\
&\cdots\cdots\cdots\cdots
\end{aligned}\right\}$$

with the coefficients c_n as unknowns. Determining the coefficients from this system we can find the values of the function $\tilde{f}(\theta)$ by applying formula (35).

Incidentally, we have proved the following assertion which is of theoretical interest: *the function $f(t)$ (and therefore the function $F(p)$) are completely specified by the values $F(p_n)$ assumed by the function $F(p)$ on an arbitrary sequence of real values $p_n \to \infty$ forming an arithmetic progression.*

§ 2. APPLICATIONS

1. Basic Idea. Suppose that we are interested in determining functions $x(t)$, $y(t)$, ... of time t which describe a certain process. A mathematical interpretation of such a problem usually reduces to solving equations connecting these and some other functions. But it often turns out that the equations connecting the transforms of the unknown functions are simpler and can be more easily solved than the equations for the preimages. For instance, as was shown in Sec. 1.3, if an unknown function is differentiated its transform is multiplied by p and if it is integrated the transform is divided by p, and therefore operations on transforms are usually simpler than the corresponding operations on the initial functions. In these cases we can pass from equations for originals to the corresponding equations for their transforms, determine the transforms and then return to the originals.

Of course, by far not all equations can be solved in this manner. First of all, for the applicability of this method the equations must be linear or reducible to linear relations. It is also desirable that the coefficients of the equations be constant because, if otherwise, the application of the method encounters many difficulties. But if these conditions are fulfilled the passage from the originals to their transforms (referred to as an *operational method*) facilitates the determination of the unknown functions of time. This method uses the system of rules and formulas (specifying the relationships between originals and their transforms) which constitute an *operational calculus*; the theory of the Laplace transformation is usually applied to establishing these rules and basic formulas.

Some complicated applications involve a number of such transitions from originals to the transforms, and the functions of time found at a certain stage in the solution of the problem may be used for determining some other functions, etc. Then it is more convenient to perform all the operations on the transforms and return to the originals we are interested in only at the final stage.

Here we shall present some examples of the application of the operational method.

2. Ordinary Differential Equations. Linear differential equations with constant coefficients and systems of such equations (e.g. see [107], Secs. XV.17, 18, 21) form a wide field of application of the operational calculus. For example, let it be required to solve the equation

$$ax''+bx'+cx=f(t)\quad\left(x=x(t),\ x'=\frac{dx}{dt},\ x''=\frac{d^2x}{dt^2}\right) \tag{1}$$

with the initial conditions

$$x(0)=\alpha,\ x'(0)=\beta \tag{2}$$

Applying the Laplace transformation to both sides of equation (1) and using its linearity (i.e. the fact that a sum of originals is transformed into the sum of their transforms, etc.; see Sec. 1.1) and Property 4 (Sec. 1.3) we obtain

$$a(p^2X-\alpha p-\beta)+b(pX-\alpha)+cX=F(p)\qquad(X=X(p))$$

where $X(p)$ is the Laplace transform of the sought-for function $x(t)$ and $F(p)$ is the transform of the given function $f(t)$. We thus receive an algebraic equation of the first degree with the function X as unknown. Writing this equation in the form

$$(ap^2+bp+c)X=F(p)+a\alpha p+a\beta+b\alpha \tag{3}$$

we find

$$X=\frac{F(p)+a\alpha p+a\beta+b\alpha}{ap^2+bp+c} \tag{4}$$

(Note that the denominator in (4) is nothing but the *characteristic polynomial* of equation (1); e.g. see [107], Sec. XV.17. This is not an accidental coincidence but a general rule.) If the function $x(t)$ is not the final result of the solution of the problem but is used, for instance, as the right-hand side of some other equation of type (1) we should consider this intermediate problem as being solved because it is the transform of the right-hand member but not the right-hand member itself that enters into a formula of type (4). But if $x(t)$ is the unknown function whose determination is required in the problem we must additionally pass from formula (4) to the original function by applying the methods of Secs. 1.4 and 1.5. The residues in formula (1.26) then correspond to the zeros of the characteristic polynomial (why?) and to the singular points of the function $F(p)$.

We see that when a differential equation is solved by means of the operational method conditions (2) are taken into account at the initial stage of the solution of the problem, which is very convenient. The calculations become particularly simple when

there are zero initial values in formulas (2) because then we can simply write

$$X(p)=\frac{F(p)}{D(p)}=Z(p)F(p) \tag{5}$$

where $D(p)$ is the characteristic polynomial and $Z(p)=\frac{1}{D(p)}$ (the latter is called the **transfer function**). As is known (e.g. see [107], the end of Sec. XV.18), the function $f(t)$ in equation (1) can often be interpreted as an external action upon a physical system (an *input-output system*) whose properties are characterized by the coefficients on the left-hand side, and the function $x(t)$ as the response of the system to this action; in other words, $f(t)$ is an **input signal (input function** or, simply, an **input)**, and $x(t)$ is the corresponding **output signal (output function** or, simply, an **output)** (see Fig. 61). Thus, for zero initial values the transform of the output function can be obtained from the transform of the input function if we simply multiply the latter by the transfer function. This makes the operational method particularly

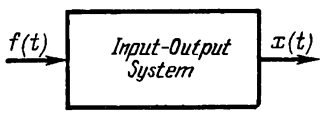


Fig. 61

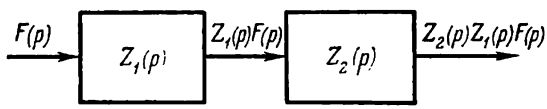


Fig. 62

convenient for investigating complicated systems composed of several input-output units in which the outputs of certain units are used as inputs for other units, etc. (Fig. 62).

If the initial values are different from zero formula (4) can be written in the form

$$X(p)=Z(p)F(p)+(a\alpha p+a\beta+b\alpha)Z(p)$$

and thus there appear additional terms in the output which are determined by the properties of the system and by the initial values but are independent of the external action.

Let us discuss the physical meaning of the inverse transform $z(t)$ of the transfer function; note that $Z(p)$ in (5) is obtained from $X(p)$ if we put $F(p)\equiv 1$, and therefore Example 5 in Sec. 1.2 indicates that $z(t)$ is the output corresponding to an external action described by the delta function $\delta(t)$, i.e. to unit pulse applied to the system at the initial moment of time on condition that the system is in the state of rest before that action. We can follow another argument to draw this conclusion:

immediately after the unit pulse has been applied the external action is switched off, and this can be interpreted as if we had $f(t) \equiv 0$ but the system gained the initial values $x(+0) = 0$ and $x'(+0) = \frac{1}{a}$ (e.g. see analogous argument in [107], the end of Sec. XV.16); then formula (4) shows that $X(p) = Z(p)$.

Passing to the inverse transforms in formula (5) we derive, on the basis of Property 6, Sec. 1.3, the relation

$$x(t) = \int_0^t z(t-\tau) f(\tau)\, d\tau \tag{6}$$

The solution in form (6) can also be obtained by means of general considerations related to the construction of the corresponding *influence function* (e.g. see [107], Sec. XIV.26). Here the influence function is of the form $z(t-\tau)$, but not $z(t, \tau)$ as in the general case, because the system in question is not only linear but also *autonomous* (i.e. its properties are independent of time), and therefore when the initial moment of time is shifted by τ the corresponding output function undergoes the same shift without changing its form.

The transfer function $Z(p)$ itself can be interpreted as follows. Suppose that the external action has the form $f(t) = e^{pt}$ where p is a fixed complex number. Then one of the particular solutions of equation (1) is $x(t) = \frac{1}{D(p)} e^{pt} = Z(p) e^{pt}$ (provided that $D(p) \neq 0$, i.e. if there is no resonance). In particular, if $p = i\omega$, i.e. p is a pure imaginary number, we have a harmonic external action, and the transfer function is thus connected with the notion of the **amplification factor** (the **gain coefficient**) for a harmonic signal; in this case $Z(i\omega)$ is referred to as the **frequency characteristic** of the physical system. If the system is stable, that is if it can only have damped free oscillations (this also means that the function $Z(p)$ has no poles other than in the region $\operatorname{Re} p < 0$), the solution of the equation (the output of the system) corresponding to the external action $e^{i\omega t}$ and any initial conditions tends, as t increases, to $Z(i\omega) e^{i\omega t}$ at the rate of an exponential function.

Sometimes it is preferable to consider the output $z_1(t)$ which corresponds not to an impulse function but to the **unit step function**

$$f(t) = \begin{cases} 0 & \text{for } t < 0 \\ 1 & \text{for } t \geqslant 0 \end{cases}$$

(also termed the **Heaviside unit function** or **unit step input**).

Then, for zero initial values, we obtain from (5) the relation

$$Z_1(p)=Z(p)\cdot\frac{1}{p},\ \text{i.e.}\ Z(p)=pZ_1(p)$$

whence it follows that for an arbitrary function $f(t)$ formula (5) yields

$$X(p)=pZ_1(p)F(p)$$

Coming back to the originals we receive

$$x(t)=\left(\int_0^t z_1(\tau)f(t-\tau)\,d\tau\right)'=z_1(t)f(0)+\int_0^t z_1(\tau)f'(t-\tau)\,d\tau$$

This formula can easily be deduced from (6) if we take into account that $z(t)=z_1'(t)$.

The operational method can be applied not only to initial-value problems but also to *boundary-value problems* (e.g. see [107], Sec. XV.16) although the calculations then become more complicated. For instance, let it be necessary to solve equation (1) on the interval $0\leqslant t\leqslant a$ for the boundary conditions $x(0)=\alpha$ and $x(T)=\gamma$. Suppose that the solution is extended as a smooth function (i.e. having a continuous derivative $x'(t)$) to the whole semiaxis $0\leqslant t<\infty$. Let the value $x'(0)$ (which is unknown) be denoted by β. Then applying to (1) the Laplace transformation we can pass to formula (4) in which β is an unknown parameter. Returning to the inverse transforms we derive, for the solution, an expression of the form $x=\varphi(t,\beta)$ after which the parameter β is found from the second boundary condition $\varphi(T,\beta)=\gamma$.

Here we also consider an example of a differential equation with variable coefficients. Take the equation

$$tx''+x'+tx=0 \qquad (7)$$

Putting $x'(0)=\alpha$ and $x(0)=\beta$, passing to the Laplace transforms and using Properties 4 and 5 from Sec. 1.3 we obtain the equation

$$-(p^2X-\beta p-\alpha)'+(pX-\beta)-X'=0,\quad \text{i.e.}\quad (p^2+1)\frac{dX}{dp}+pX=0$$

Thus, for $X(p)$ we have an ordinary differential equation of the first order. Solving the equation we find

$$X(p)=\frac{C}{\sqrt{p^2+1}}$$

whence

$$x(t)=CJ_0(t)$$

(see formula (1.29)). Equation (7) is a special case of the Bessel equation for $p=0$ (e.g. see [107], equation (XV.167)). As is known, there are two linearly independent particular solutions of the Bessel equation but we have found only one of them because

the other is unbounded in the vicinity of the point $t=0$, and therefore Property 4 from Sec. 1.3 does not apply to it.

If, instead of equation (1), we take a system of linear differential equations with constant coefficients, then, after the Laplace transformation has been applied to the equations, we obtain a *matrix equation* (e.g. on such equations see [107], Sec. XI.3) of the form

$$D(p)\,X(p)=F(p) \tag{8}$$

where $D(p)$ is the *characteristic matrix* of the system, $X(p)$ is the column vector of the transforms of the sought-for functions and $F(p)$ is the column vector composed of the transforms of the given functions (more precisely, for zero initial values the components of this column vector are the transforms of the nonhomogeneity terms of the system and for nonzero values they additionally include some terms analogous to those in the solution of equation (3)). Solving equation (8) we obtain

$$X(p)=Z(p)\,F(p)$$

where $Z(p)=[D(p)]^{-1}$ is the **transfer matrix**. According to the rule of constructing an inverse matrix (e.g. see [107], Sec. XI.3), its elements have the characteristic polynomial $\det D(p)$ as denominator, and therefore, when passing to the originals, we must take into account the poles corresponding to the zeros of the polynomial (i.e. to the roots of the characteristic equation).

3. Difference and Difference-Differential Equations. We begin with *difference equations* (e.g. see [107], the end of Sec. XVII.16). For example, take the equation

$$ax(t)+bx(t-h)+cx(t-k)=f(t) \qquad (0\leqslant t<\infty) \tag{9}$$

where $0<h<k$. We can set an initial condition by specifying the values of $x(t)$ for $-k\leqslant t<0$. Then, at least in principle, we can use the so-called **step-by-step computation** method: when t varies from 0 to h the values of the second and third terms on the left-hand side of (9) are known (why?), and therefore it is possible to determine the solution $x(t)$ for $0\leqslant t<h$; then we pass to the values of t ranging from h to $2h$ for which the values of $x(t-h)$ and $x(t-k)$ have already been found (why?) and determine from (9) $x(t)$ for $h\leqslant t<2h$, etc. This method can be applied to the numerical computation of the solution even when the equation is nonlinear and nonautonomous but it is by far not so convenient for a theoretical investigation of the solution.

To solve a linear autonomous system of difference equations we can use the operational calculus. To illustrate this application let us again take equation (9) and consider the Laplace transforms

of its both sides taking into account Property 2, Sec. 1.3. Equating these transforms we obtain

$$aX+be^{-ph}X+be^{-ph}\int_{-h}^{0}e^{-pt}x(t)\,dt+ce^{-pk}X+$$

$$+ce^{-pk}\int_{-k}^{0}e^{-pt}x(t)\,dt=F(p) \qquad (10)$$

whence $X(p)$ is readily found (note that the integral terms only involve the values of $x(t)$ for $t<0$ which are set beforehand in accordance with the initial conditions). As in Sec. 1, the case of a zero initial value is the simplest one because formula (10) then implies a formula of type (5) in which the transfer function is

$$Z(p)=\frac{1}{a+be^{-ph}+ce^{-pk}} \qquad (11)$$

The poles of this function (at which we must take the residues) are determined in a simple way when the numbers h and k are *commensurable* (in the sense that their quotient is a rational number), i.e. when $h=m\tau$ and $k=n\tau$, m and n being integers. Then, in order to determine the zeros of the denominator, we must solve the algebraic equation

$$a+bu^m+cu^n=0 \qquad (u=e^{-p\tau})$$

and then put $p=-\frac{1}{\tau}\operatorname{Ln} u$. If h and k are *incommensurable* we obtain a transcendental characteristic equation which is to be solved numerically.

So-called **difference-differential equations** can be investigated in an analogous manner. For instance, take the equation

$$ax'(t)+bx(t)+cx'(t-h_1)+dx(t-h_2)=f(t) \qquad (0\leqslant t<\infty) \quad (12)$$

belonging to this type in which $h_1>0$ and $h_2>0$. An initial condition can be imposed if we specify the values of $x(t)$ for $-h\leqslant t<0$ where h is the greatest of the numbers h_1 and h_2. Besides, the value $x(+0)$ is also set in advance (it may be different from $x(-0)$). Here we can also apply the step-by-step computation, the length of each step being equal to the least of the numbers h_1 and h_2. Let the reader determine the Laplace transform of the solution and show that in this case the transfer function is

$$Z(p)=(ap+b+cpe^{-h_1p}+de^{-h_2p})^{-1}$$

The determination of its poles reduces to computing the roots of a *quasi-polynomial* (see formula (II.3.46)).

It should be noted that when we consider an initial-value problem for an equation of type (12) the equation must contain a term with the highest derivative which is taken for the greatest value of the argument. (For instance, we cannot add a term involving $x''(t-h_1)$ to equation (12)). Indeed, if otherwise, it can be proved that the *characteristic quasi-polynomial* has zeros with arbitrarily large real parts and therefore the operational method is inapplicable to the solution of the problem (why?).

4. Integral and Integro-Differential Equations. The operational calculus can be applied to many *integral equations* (which involve an unknown function under the integral sign) and also to *integro-differential equations* (i.e. ones containing some sought-for functions both under the sign of integration and under the sign of differentiation). Here we do not present the general theory of such equations and only consider two simple examples.

1. Take the integral equation

$$\int_0^t x(\tau)\, g(t-\tau)\, d\tau = f(t) \qquad (0 \leqslant t < \infty)$$

where the functions f and g are known and x is the sought-for function. Using Property 6 from Sec. 1.3 we find that $X(p)\,G(p) = F(p)$, that is

$$X(p) = \frac{F(p)}{G(p)} \tag{13}$$

By the way, here we must additionally check whether the right-hand side can serve as a Laplace transform (see Sec. 1.4), which is by far not always so for arbitrary functions f and g since the right-hand member does not necessarily tend to zero for $\operatorname{Re} p \geqslant s$, $|p| \longrightarrow \infty$.

There are some nonlinear integral equations that are also solvable by means of the operational method but this only refers to equations of a special type. For instance, applying the Laplace transformation to the equation

$$\int_0^t x(\tau)\, x(t-\tau)\, d\tau = f(t) \qquad (0 \leqslant t < \infty)$$

we obtain the equation $X^2(p) = F(p)$ for the transforms whence we find $X(p) = \sqrt{F(p)}$, etc.

2. Let us consider the integro-differential equation

$$x'(t) + \int_0^t x(\tau)\, g(t-\tau)\, d\tau = f(t) \qquad (0 \leqslant t < \infty)$$

with the additional condition $x(+0)=\alpha$ (where the value of α is given). Passing to the transforms we obtain

$$pX(p)-\alpha+X(p)\,G(p)=F(p)$$

which yields

$$X(p)=\frac{F(p)+\alpha}{p+G(p)}$$

In this case the condition that the right member must tend to zero for $p\to\infty$ is fulfilled (why?), and hence we have in fact found the Laplace transform of the solution.

5. Partial Differential Equations. The theory of *partial differential equations* (e.g. see [107], Secs. XVII.31, 34 and Appendix) is an extensively developed branch of mathematics but here we shall not deal with it and shall only give an example of the application of the operational calculus to solving the so-called **heat conduction equation**

$$\frac{\partial u}{\partial t}=a\frac{\partial^2 u}{\partial x^2} \tag{14}$$

In courses of equations of mathematical physics it is proved that this equation describes the distribution of temperature $u=u(t, x)$, at time moment t at the point x of a rectilinear rod when heat can only be transferred through its ends. The parameter a (termed the **coefficient of temperature conductivity**) is specified solely by the properties of the substance of the rod.

We begin with the case of an infinite rod, i.e. $-\infty<x<\infty$. Let the distribution of temperature at the initial moment of time be described by the relation

$$u|_{t=0}=\varphi(x) \tag{15}$$

where the function $\varphi(x)$ is known. Thus, the mathematical problem is reduced to solving equation (14) for the given initial condition (15). To find the solution we apply the Laplace transformation to both members of equation (14) considering x to be a parameter, that is $u(t, x)\to U(p, x)$. Using the rule of differentiation with respect to the parameter (see the end of Sec. 1.1) and formula (1.9) we obtain the equation

$$pU-\varphi(x)=a\frac{\partial^2 U}{\partial x^2}$$

For every fixed value of p this is an ordinary differential equation which can be rewritten in the form

$$a\frac{d^2U}{dx^2}-pU=-\varphi(x) \tag{16}$$

It is the basic idea of the application of the Laplace transformation to partial differential equations that one of the indepen-

dent variables becomes a parameter with respect to which the transformation is performed and thus the number of the arguments is reduced by unity. Naturally, the smaller the number of the independent variables, the simpler the resultant equation; in particular, in the case of two independent variables the application of the Laplace transformation leads to one independent variable in the transformed equation, that is we arrive at an ordinary differential equation (which is the case in this example).

In solving equation (16) we begin with the case $\varphi(x)=\delta(x)$ (the delta function). Then, for $x\neq 0$, equation (16) becomes homogeneous and therefore has the general solution

$$U=C_1e^{\sqrt{\frac{p}{a}}x}+C_2e^{-\sqrt{\frac{p}{a}}x}$$

where the arbitrary constants C_1 and C_2 may assume different values for $x<0$ and $x>0$; besides, they may depend on p (they only are constant as functions of x). Let us agree that for large values of $\operatorname{Re} p$ the expression $\sqrt{\frac{p}{a}}$ is understood as the branch of the square root which takes on positive values for real $p>0$. Then the condition that $U(p,\ x)$ turns into zero for $p\longrightarrow\infty$ (this is the general property of the Laplace transforms) implies that

$$U(p,\ x)=\begin{cases} C_1(p)e^{\sqrt{\frac{p}{a}}x} & \text{for } x<0 \\ C_2(p)e^{-\sqrt{\frac{p}{a}}x} & \text{for } x>0 \end{cases} \tag{17}$$

But if we put $\varphi(x)=\delta(x)$ in (16) and integrate it twice from -0 to $+0$ we arrive at the equalities

$$a\left(\frac{dU}{dx}\Big|_{x=+0}-\frac{dU}{dx}\Big|_{x=-0}\right)=-1,\ U|_{x=+0}-U|_{x=-0}=0$$

It follows that we have $C_1=C_2=\frac{1}{2\sqrt{ap}}$ in (17) (check it up!), that is

$$U(p,\ x)=\frac{1}{2\sqrt{ap}}e^{-\sqrt{\frac{p}{a}}|x|} \qquad (x\gtrless 0) \tag{18}$$

We see that the right-hand side of (18) is obtained from the right-hand member of the tabular formula (1.25) by substituting $\frac{x^2}{a}p$ for p and multiplying the result by $\frac{|x|}{2a}$. Hence, it follows from formula (1.25) and Property 1, Sec. 1.3, that the inverse

transform of function (18) is equal to

$$u(t, x)=\frac{|x|}{2a}\frac{a}{x^2}\frac{1}{\sqrt{\pi\left(\frac{a}{x^2}t\right)}}e^{-\frac{1}{4\frac{a}{x^2}t}}=\frac{1}{2\sqrt{\pi at}}e^{-\frac{x^2}{4at}} \qquad (19)$$

(check it up!)

If we now put $\varphi(x)=\delta(x-\xi)$, the invariance of equation (14) with respect to the shifts along the x-axis implies that the corresponding solution is also obtained from (19) by substituting $x-\xi$ for x. It follows that if $\varphi(x)$ is an arbitrary function, the application of the *superposition principle* (e.g. see [107], Sec. XIV.26) yields the final solution

$$u(t, x)=\int_{-\infty}^{\infty}\frac{1}{2\sqrt{\pi at}}e^{-\frac{(x-\xi)^2}{4at}}\varphi(\xi)\,d\xi$$

of the problem which for the first time was obtained by S. D. Poisson by means of another method.

Now we shall consider another problem connected with equation (14), namely the case when there is a heat transfer through the terminal of a semi-infinite rod on condition that the law of variation of temperature at that end-point is known, and the temperature of the rod at the initial time moment is identically equal to zero. The corresponding mathematical problem reduces to solving equation (14) for $0\leqslant t<\infty$, $0\leqslant x<\infty$, with the initial condition

$$u|_{t=0}=0 \qquad (0\leqslant x<\infty)$$

and the boundary condition

$$u|_{x=0}=\varphi(t) \qquad (20)$$

where $\varphi(t)$ is a given function. Let us apply the Laplace transformation (with respect to t) to equation (14) taking into account condition (20). This results in

$$pU(p, x)=a\frac{\partial^2 U}{\partial x^2}, \qquad U(p, 0)=\Phi(p)$$

By analogy with (17), we obtain the solution of the latter problem in the form

$$U(p, x)=\Phi(p)e^{-\sqrt{\frac{p}{a}}x} \qquad (21)$$

According to (1.24), the inverse transform of the function $e^{-\sqrt{\frac{p}{a}}x}$ is

$$\frac{a}{x^2}\frac{1}{2\sqrt{\pi}}\left(\frac{a}{x^2}t\right)^{-\frac{3}{2}}e^{-1/[4(at/x^2)]}=\frac{x}{2\sqrt{\pi a}\,t^{\frac{3}{2}}}e^{-\frac{x^2}{4at}}$$

and therefore, by Property 6, Sec. 1.3, the inverse transform of function (21) is expressed by the formula

$$u(t,\ x)=\frac{x}{2\sqrt{\pi a}}\int_0^t(t-\tau)^{-\frac{3}{2}}e^{-\frac{x^2}{4a(t-\tau)}}\varphi(\tau)\,d\tau$$

which thus gives the solution of the problem in question.

We see that the application of the operational method is not only connected with the basic properties of the Laplace transformation but also involves various formulas determining the transition from concrete initial functions to their transforms and vice versa.

Here we also give an example leading to a sum of an infinite series. Let us consider a finite rod and solve equation (14) for $0\leqslant x\leqslant l$, $0\leqslant t<\infty$ with the initial condition

$$u|_{t=0}=0 \qquad (0\leqslant x\leqslant l)$$

(which expresses the fact that at the initial time moment the temperature is identically equal to zero) and the boundary conditions

$$u|_{x=0}=u_0,\quad u|_{x=l}=0 \qquad (0\leqslant t<\infty) \tag{22}$$

(which mean that the temperature u_0 at the left end-point is kept constant and the temperature at the right end-point is equal to zero). Applying to equation (14) with conditions (22) the Laplace transformation we obtain

$$pU=a\frac{\partial^2U}{\partial x^2},\quad U|_{x=0}=\frac{u_0}{p},\quad U|_{x=l}=0$$

Let the reader solve the latter boundary-value problem and show that for any fixed value of the parameter p its solution is given by the formula

$$U(p,\ x)=\frac{u_0}{p}\frac{\sinh\sqrt{\frac{p}{a}}(l-x)}{\sinh\sqrt{\frac{p}{a}}\,l}$$

For every fixed value of x this function is one-valued (why?) and

analytic in the whole p-plane except the point $p_0=0$ and the points determined by the equation $\sinh \sqrt{\frac{p}{a}}\, l=0$ whence

$$e^{2\sqrt{\frac{p}{a}}\, l}=1, \quad 2\sqrt{\frac{p}{a}}\, l=2k\pi i, \quad p_k=-\frac{ak^2\pi^2}{l^2} \qquad (k=1,\ 2,\ \ldots)$$

At these points the function U has poles of the first order. Using formula (1.26) (but applying it to an infinite number of summands) and formula (II.3.5) we find

$$u(t,\ x)=u_0\left[\frac{l-x}{l}-\sum_{k=1}^{\infty}\frac{l^2}{ak^2\pi^2}\,\frac{\sinh\left(i\frac{k\pi}{l}(l-x)\right)}{\cosh\left(i\frac{k\pi}{l}l\right)\frac{1}{2}\frac{l}{iak\pi}l}\,e^{-\frac{ak^2\pi^2}{l^2}t}\right]=$$

$$=u_0\left(1-\frac{x}{l}-\frac{2}{\pi}\sum_{k=1}^{\infty}\frac{1}{k}\sin\frac{k\pi x}{l}e^{-\frac{ak^2\pi^2}{l^2}t}\right) \qquad (23)$$

Let us consider the state of the rod after the process becomes stationary (that is time-independent, which, formally, is achieved when $t\longrightarrow\infty$). The limiting formula (1.32) implies that

$$u(\infty,\ x)=\lim_{p\to 0}[pU(p,\ x)]=\lim_{p\to 0}u_0\frac{\sinh\sqrt{\frac{p}{a}}(l-x)}{\sinh\sqrt{\frac{p}{a}}\,l}=u_0\frac{l-x}{l}$$

By the way, the same result also follows from formula (23) describing the transient process.

In conclusion we note that the operational calculus has many important mathematical applications to computing definite integrals, deriving various formulas connecting special functions, etc; e.g. see [5] and [29].

§ 3. SOME GENERALIZATIONS

There are many other *integral transformations* similar to the Laplace and Fourier transformations which are applied to various classes of problems. These applications follow the general scheme given in § 2: the problem in question is reformulated in terms of the transforms of the unknown functions, and if the reduced problem can be solved by means of some simple techniques the inverse transformation is applied to its solutions, which yields the solution of the original problem. Here we shall give a brief review of such transformations referring the reader for more detail to special monographs.

1. Discrete Laplace Transformation. The discrete Laplace transformation associates with a given number sequence x_0, x_1, x_2, ... (we shall denote it by $\{x_n\}$) the corresponding function $X(p)$ of the complex variable $p=s+ir$ determined by the formula

$$\{x_n\} \longrightarrow X(p) = \sum_{n=0}^{\infty} x_n e^{-np} \tag{1}$$

This transformation is widely applied in the theory of impulse systems. If, for $n \longrightarrow \infty$, the variable x_n remains bounded or increases, in its modulus, at exponential rate of growth, the function $X(p)$ is analytic for sufficiently large values of s and is periodic with respect to r with period 2π. We readily prove that

$$\left.\begin{array}{l} \{1, 0, 0, \ldots\} \longrightarrow 1, \ \{1, 1, 1, \ldots\} \longrightarrow \dfrac{e^p}{e^p-1} \\ \{e^{an}\} \longrightarrow \dfrac{e^p}{e^p-e^a}, \ \{n^k e^{an}\} \longrightarrow \dfrac{k! e^p}{(e^p-e^a)^{k+1}} \end{array}\right\} \tag{2}$$

($k=0, 1, 2, \ldots$), etc. The properties of this transformation are similar to those described in Sec. 1.3 (for more detail we refer the reader to [30] and [34]). For instance,

$$\{x_{n+1}\} \longrightarrow \sum_{n=0}^{\infty} x_{n+1} e^{-np} = |n+1=m| = \sum_{m=1}^{\infty} x_m e^{-(m-1)p} =$$

$$= e^p \left(\sum_{m=0}^{\infty} x_m e^{-mp} - x_0 \right) = e^p X(p) - x_0 e^p \tag{3}$$

In particular, the discrete transformation can be applied to solving linear difference equations with constant coefficients (e.g. see [107], Sec. XVII.16). For example, let it be required to solve the equation

$$\alpha x_n + \beta x_{n+1} + \gamma x_{n+2} = 0 \qquad (n=0, 1, 2, \ldots) \tag{4}$$

(e.g. see [107], equation (XVII.68)) in which the numbers x_0 and x_1 are given. Performing the discrete Laplace transformation in equation (4), using Property (3) and taking into account linearity we get

$$\alpha X(p) = \beta (e^p X(p) - x_0 e^p) + \gamma (e^{2p} X(p) - x_0 e^{2p} - x_1 e^p) = 0$$

whence

$$X(p) = \frac{\gamma x_0 e^{2p} + (\beta x_0 + \gamma x_1) e^p}{\gamma e^{2p} + \beta e^p + \alpha} \tag{5}$$

Now we can expand expression (5), regarded as a rational fraction with respect to e^p, into partial rational fractions (e.g.

see [107], Sec. VIII.10) and use formulas (2) taking them in the form

$$\frac{1}{e^p - e^a} \leftarrow e^{-a}\{e^{an}\} - \{e^{-a},\ 0,\ 0,\ \ldots\}$$

etc. We can also apply the inversion formula given below.

Let us write transformation (1) in the form

$$X(s+ir) = \sum_{n=-\infty}^{\infty} (x_n e^{-ns})\, e^{-inr}$$

where we put $x_n = 0$ for $n < 0$. Then the inversion formula immediately follows from the formula for the coefficients of Fourier's series in complex form (e. g. see [107], Sec. XVII.26): we have

$$x_n e^{-ns} = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(s+ir)\, e^{inr}\, dr$$

which implies

$$x_n = \frac{1}{2\pi i} \int_{s-i\pi}^{s+i\pi} X(p)\, e^{np}\, dp \qquad (n = 0,\ 1,\ 2,\ \ldots) \tag{6}$$

where s is an arbitrary sufficiently large number. This proof also shows that for $n = -1,\ -2,\ \ldots$ integral (6) is equal to zero.

This transformation is sometimes taken in another form which is obtained by putting $e^p = z$. Then the transform of an arbitrary sequence $\{x_n\}$ is the function $\tilde{X}(z)$ determined by the relation

$$\tilde{X}(z) = X(p) = \sum_{n=0}^{\infty} x_n z^{-n} \tag{7}$$

Since the values of $\mathrm{Re}\, p$ must be sufficiently large while those of $\mathrm{Im}\, p$ are arbitrary, series (7) is convergent for all values of z with sufficiently large moduli, and hence it is Laurent's expansion in the neighbourhood of the point at infinity of the function $\tilde{X}(z)$ which is analytic at $z = \infty$ (see Sec. II.3.6). Formulas (2) and the properties of the discrete transformation taken in this form are accordingly changed $\left(\text{for instance, } \{k^n\} \longrightarrow \frac{z}{z-k}, \text{ etc.}\right)$. Inversion formula (6) can be rewritten as

$$x_n = \frac{1}{2\pi i} \oint_{|z|=R} \tilde{X}(z)\, z^{n-1}\, dz \qquad (n = 0,\ 1,\ \ldots)$$

where R is sufficiently large and the circle $|z| = R$ is positively oriented. The computation of this integral can be carried out by means of the theory of residues (let the reader apply this technique to Example 5).

In conclusion we note that the result of the discrete transformation of a sequence $\{x_n\}$ is nothing but the ordinary Laplace transform of the *impulse function*

$$\sum_{n=0}^{\infty} x_n \delta(t-n)$$

(check it up!).

2. Fourier Transformation of Functions of Bounded Order of Growth. The application of Fourier's transformation is limited by the requirement that the original function should be absolutely integrable and must therefore vanish at infinity. But if we consider Fourier's transform as an analytic function this restriction can be discarded. Let us begin with a function $f(x)$ which increases, for $t \to \infty$, at exponential rate of growth, and vanishes for $x < x_0$:

$$f(x) \equiv 0 \text{ for } -\infty < x < x_0, \quad |f(x)| \leqslant Me^{\alpha x} \text{ for } x_0 < x < \infty \qquad (8)$$

(In this case we say that $f(x)$ has a **bounded order of growth** as $t \to \infty$, and the greatest lower bound of those values of α for which inequality (8) holds is called the **index of growth** of the function $f(x)$.) We shall consider the variable k entering into the formula

$$\hat{f}(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x) e^{-ikx} dx \qquad (9)$$

(determining the *Fourier transform*) to be complex. Then it can readily be shown that $\hat{f}(k)$ is an analytic function of k in a region which contains the half-plane $\operatorname{Im} k < -\alpha$. Now putting $ik = p$ and $\hat{f}(k) = F(p)$ we arrive (for $x_0 \geqslant 0$) at the Laplace transform (1.1) and hence all the properties of the Laplace transformation studied in the foregoing sections can be reformulated for transformation (9). In particular, the inversion formula turns into

$$f(x) = \int_{-x+is}^{x+is} \hat{f}(k) e^{ikx} dk \qquad (\operatorname{Im} k = s < -\alpha)$$

Now let us consider the case when the function $f(x)$ also increases for $x \to -\infty$. Then we can apply the following technique. Suppose that

$$|f(x)| \leqslant M_+ e^{\alpha_+ x} \text{ for } 0 < x < \infty \text{ and } |f(x)| \leqslant M_- e^{\alpha_- x} \text{ for } -\infty < x < 0$$

Let us put

$$f_+(x) = \begin{cases} 0 \text{ for } -\infty < x < 0 \\ f(x) \text{ for } 0 < x < \infty \end{cases} \text{ and } f_-(x) = \begin{cases} f(x) \text{ for } -\infty < x < 0 \\ 0 \text{ for } 0 < x < \infty \end{cases}$$

Then the corresponding Fourier transforms $\hat{f}_+(k)$ and $\hat{f}_-(k)$ are

analytic functions of k: the former for $\operatorname{Im} k < -\alpha_+$ and the latter for $\operatorname{Im} k > -\alpha_-$. The inversion formula can thus be written in the form

$$f(x) = \int_{-\infty+is_+}^{\infty+is_+} \hat{f}_+(k)\, e^{ikx}\, dk + \int_{-\infty+is_-}^{\infty+is_-} \hat{f}_-(k)\, e^{ikx}\, dk \qquad (10)$$

where $s_+ < -\alpha_+$ and $s_- > -\alpha_-$.

It can readily be shown that if we replace the constant M in inequality (8) by an integrable (summable) positive function $M(x)$ integral (9) satisfies, for $\operatorname{Im} k \leqslant -\alpha$, the conditions for continuous dependence on the parameter (e. g. see [107], Sec. XIV.21) and therefore the function $\hat{f}(k)$ is continuous in the closed half-plane $\operatorname{Im} k \leqslant -\alpha$ including the straight line bounding it. The continuity properties of the functions $\hat{f}_+$ and $\hat{f}_-$ are formulated in like manner. In particular, if the function $f(x)$ is absolutely integrable we can put $\alpha_+ = \alpha_- = 0$. Then we have the following situation: the function $\hat{f}_+(k)$ is analytic in the lower half-plane of the variable k and is continuous on the real axis bounding this half-plane (but its analyticity can be violated at the points of the axis), the function $\hat{f}_-(k)$ is analytic in the upper half-plane and also continuous on the real axis (but here the variable point k can only approach the real axis from below, in contrast to the former case), and at the points of the real axis we have $\hat{f}_+(k) + \hat{f}_-(k) = f(k)$. (The latter assertion obviously follows from equality (10) if we pass to the limit for $s_+ \to 0$ and $s_- \to 0$.)

We shall demonstrate the application of the above notions to deriving **Poisson's summation formula** for the series

$$S = \sum_{n=-\infty}^{\infty} f(\beta n) \qquad (\beta > 0)$$

where $f(x)$ $(-\infty < x < \infty)$ is a continuous summable function. For this purpose we shall use formula (10) in which s_+ and s_- are arbitrary positive numbers. Let us substitute $x = \beta n$ into (10) and perform the summation with respect to n from $-\infty$ to ∞. The first integral, which is equal to $f_+(x)$, turns into zero for $n < 0$, and consequently we must only carry out the summation for $n \geqslant 0$; the second integral possesses analogous properties, and therefore we obtain

$$S = \sum_{n=0}^{\infty} \int_{-\infty+is_-}^{\infty+is_+} \hat{f}_+(k)\, e^{ik\beta n}\, dk + \sum_{n=0}^{-\infty} \int_{-\infty+is_-}^{\infty+is_-} \hat{f}_-(k)\, e^{ik\beta n}\, dk =$$

$$= \int_{-\infty+is_+}^{\infty+is_+} \hat{f}_+(k)\,(1 - e^{i\beta k})^{-1}\, dk + \int_{-\infty+is_-}^{\infty+is_-} \hat{f}_-(k)\,(1 - e^{-i\beta k})^{-1}\, dk \qquad (11)$$

Let us assume, for simplicity's sake, that the function $f(x)$ tends to zero for $x \to \pm\infty$ and has an exponential order of smallness. (It turns out in fact that this condition may be rejected.) Then we can take $s_+ > 0$, $s_- < 0$ and add, respectively, to the contours of integration of the first and second integrals on the right-hand side of (11) the lower and upper semicircular arcs of an infinitely large radius. The integrals taken over the arcs tend to zero as the radius increases infinitely, and therefore we can apply Cauchy's residue theorem and write

$$S = -2\pi i \sum_{n=-\infty}^{\infty} \frac{1}{-i\beta} \hat{f}_+\left(\frac{2\pi n}{\beta}\right) + 2\pi i \sum_{n=-\infty}^{\infty} \frac{1}{i\beta} \hat{f}_-\left(\frac{2\pi n}{\beta}\right) =$$

$$= \frac{2\pi}{\beta} \sum_{n=-\infty}^{\infty} \hat{f}\left(\frac{2\pi n}{\beta}\right)$$

The formula

$$\sum_{n=-\infty}^{\infty} f(\beta n) = \frac{2\pi}{\beta} \sum_{n=-\infty}^{\infty} \hat{f}\left(\frac{2\pi n}{\beta}\right)$$

which thus have been derived is used in many theoretical investigations and practical computations. Let the reader obtain this result by applying the formula $\int f g^* dx = 2\pi \int \hat{f} \hat{g}^* dk$ which follows from Parseval's relation (e. g. see [107], Sec. XVII.33), choosing the function $g(x)$ in an appropriate manner and taking advantage of the formula

$$\sum_{n=-\infty}^{\infty} e^{i\beta n x} = \frac{2\pi}{\beta} \sum_{n=-\infty}^{\infty} \delta\left(x - \frac{2\pi n}{\beta}\right)$$

3. Some Other Integral Transformations. Every *integral transformation* is determined by a formula of the type

$$F(p) = \int_a^b K(p, t) f(t)\, dt \tag{12}$$

where $K(p, t)$ is a **kernel** specifying the transformation. For instance, for Fourier's transformation we have

$$a = -\infty, \quad b = \infty, \quad K(p, t) = \frac{1}{2\pi} e^{-ipt}$$

and for Laplace's transformation

$$a = 0, \quad b = \infty, \quad K(p, t) = e^{-pt}$$

Similarly, for Fourier's cosine and sine transformations (e. g. see

[107], Sec. XVII.32) we have $a=0$, $b=\infty$ and, respectively, $K(p, t)=\frac{1}{\pi}\cos pt$ and $K(p, t)=\frac{1}{\pi}\sin pt$.

Other integral transformations are in a certain way connected with Fourier's transformation by means of certain relationships. For instance, let us consider **Mellin's transformation** determined by the formula

$$F(p)=\int_0^\infty t^{p-1}f(t)\,dt \tag{13}$$

where $p=s+ir$ is a complex variable. The change of variable $t=e^{-x}$ shows (check it up!) that for every (fixed) admissible value of s the function $F(s+ir)$ is the Fourier transform of the function $2\pi e^{-sx}f(e^{-x})$. Therefore, applying the inversion formula for Fourier's transformation, we can write

$$2\pi e^{-sx}f(e^{-x})=\int_{-\infty}^{\infty}F(s+ir)\,e^{irx}\,dr$$

whence we derive the inversion formula for transformation (13):

$$f(t)=\frac{1}{2\pi i}\int_{s-i\infty}^{s+i\infty}t^{-p}F(p)\,dp \tag{14}$$

Let the reader establish the conditions on the function $f(t)$ guaranteeing the existence of the transform determined by formula (13) in a domain in the p-plane and find out the admissible values of s in formula (14). For the properties and applications of Mellin's transformation and other integral transformations we refer the reader to [28] and [129].

The **Hilbert transformation** is widely applied in the theory of *singular integral equations* (see § VII.5). Suppose that $f(t)$ is a real absolutely integrable function on the entire t-axis. Let $\hat{f}(\omega)$ be its Fourier transform and the functions $\tilde{f}(\omega)$ and $F(t)$ be determined by the formulas

$$\tilde{f}(\omega)=\begin{cases}\hat{f}(\omega) & (\omega>0),\\ 0 & (\omega<0),\end{cases}\qquad F(t)=-2\,\mathrm{Im}\int_0^\infty e^{i\omega t}\hat{f}(\omega)\,d\omega \tag{15}$$

Now note that the fact that the function $f(t)$ is real is equivalent to the property

$$\hat{f}(-\omega)=[\hat{f}(\omega)]^*$$

of its Fourier transform (prove this assertion!). Therefore we have

$$2\operatorname{Re}\int_0^\infty e^{i\omega t}\hat{f}(\omega)\,d\omega=\int_0^\infty e^{i\omega t}\hat{f}(\omega)\,d\omega+\int_0^\infty e^{-i\omega t}[\hat{f}(\omega)]^*\,d\omega=$$

$$=\int_0^\infty e^{i\omega t}\hat{f}(\omega)\,d\omega+\int_{-\infty}^0 e^{i\omega_1 t}\hat{f}(\omega_1)\,d\omega_1=\int_{-\infty}^\infty e^{i\omega t}\hat{f}(\omega)\,d\omega=f(t)$$

(where $\omega=-\omega_1$). We see that the inverse Fourier transform of the function $\tilde{f}(\omega)$ is equal to

$$\int_{-\infty}^\infty e^{i\omega t}\tilde{f}(\omega)\,d\omega=\int_0^\infty e^{i\omega t}\hat{f}(\omega)\,d\omega=\operatorname{Re}\int_0^\infty+i\operatorname{Im}\int_0^\infty=\frac{1}{2}[f(t)-iF(t)]$$

that is

$$\tilde{f}(\omega)=\frac{1}{2}[\hat{f}(\omega)-i\hat{F}(\omega)]$$

Therefore, for $\omega>0$, we can write

$$\hat{F}(\omega)=[\hat{F}(-\omega)]^*=[2i\tilde{f}(-\omega)-i\hat{f}(-\omega)]^*=0+i\hat{f}(\omega)$$

and

$$2\operatorname{Im}\int_0^\infty e^{i\omega t}\hat{F}(\omega)\,d\omega=2\operatorname{Im}\int_0^\infty e^{i\omega t}i\hat{f}(\omega)\,d\omega=2\operatorname{Re}\int_0^\infty e^{i\omega t}\hat{f}(\omega)\,d\omega=f(t) \tag{16}$$

Changing the order of integration in the second formula (15) we deduce

$$F(t)=-2\operatorname{Im}\int_0^\infty e^{i\omega t}\,d\omega\frac{1}{2\pi}\int_{-\infty}^\infty e^{-i\omega\tau}f(\tau)\,d\tau=$$

$$=-\frac{1}{\pi}\lim_{N\to\infty}\operatorname{Im}\int_{-\infty}^\infty f(\tau)\,d\tau\int_0^N e^{i\omega(t-\tau)}\,d\omega=$$

$$=\frac{1}{\pi}\lim_{N\to\infty}\int_{-\infty}^\infty f(\tau)\frac{1-\cos N(\tau-t)}{\tau-t}\,d\tau$$

The last integral can be broken up into three terms according to the scheme

$$\int_{-\infty}^\infty=\int_{-\infty}^{t-\varepsilon}+\int_{t-\varepsilon}^{t+\varepsilon}+\int_{t+\varepsilon}^\infty \tag{17}$$

where $\varepsilon>0$ is a fixed sufficiently small number. Now we see that for $N\to\infty$ the integrals corresponding to the first and third terms

on the right-hand side of (17) tend, respectively, to

$$\int_{-\infty}^{t-\varepsilon} \frac{f(\tau)}{\tau-t}\,d\tau \quad \text{and} \quad \int_{t+\varepsilon}^{\infty} \frac{f(\tau)}{\tau-t}\,d\tau$$

because the frequency of oscillation of the cosine tends to infinity (this assertion can be rigorously proved by means of the substitution $N(\tau-t)=s$). Let us consider the second integral. Expanding $f(\tau)$ into the series in powers of $\tau-t$ we see that the integral tends to zero for $\varepsilon \to 0$. Therefore, according to the definition of Cauchy's principal value of a singular (divergent) integral (e.g. see [107], Sec. XIV.19), we can finally write

$$F(t)=\frac{1}{\pi}\,\text{v.p.}\int_{-\infty}^{\infty} \frac{f(\tau)}{\tau-t}\,d\tau$$

Similarly, from formula (16) we obtain the relation

$$f(t)=-\frac{1}{\pi}\,\text{v.p.}\int_{-\infty}^{\infty} \frac{F(\tau)}{\tau-t}\,d\tau$$

These two formulas determine *Hilbert's transformation* and its inverse; they specify linear relationships and therefore are also valid for complex functions of a real argument. (Let the reader prove that for this transformation we have $\sin t \to \cos t$; in this connection also see formulas (II.3.33).)

Some integral transformations are obtained from the formulas of the ***n*-dimensional Fourier transformation** for functions of several variables. For instance, let us consider the formulas for **Fourier's two-dimensional transformation:**

$$F(\xi,\ \eta)=\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i(\xi x+\eta y)} f(x,\ y)\,dx\,dy \tag{18}$$

$$f(x,\ y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i(\xi x+\eta y)} F(\xi,\ \eta)\,d\xi\,d\eta \tag{19}$$

These formulas are nothing but the result of the successive application of ordinary (one-dimensional) Fourier's transformation with respect to each independent variable. The general formulas for the n-dimensional Fourier transformation are written in like manner.

Let us now suppose that after we have passed to polar coordinates ρ, φ the function $f(x,\ y)$ turns out to be independent of φ, i.e. $f=f(\rho)$. Then, passing to the polar coordinates σ, ψ in the

$\xi\eta$-plane, we rewrite the right-hand side of (18) in the form

$$\frac{1}{(2\pi)^2}\int_0^\infty \rho\,d\rho \int_0^{2\pi} e^{-i\,(\sigma\cos\psi\cdot\rho\cos\varphi+\sigma\sin\psi\cdot\rho\sin\varphi)} f(\rho)\,d\varphi =$$

$$= \frac{1}{(2\pi)^2}\int_0^\infty f(\rho)\,\rho\,d\rho \int_0^{2\pi} e^{-i\sigma\rho\cos(\varphi-\psi)}\,d\varphi$$

The cosine being periodic, the integral with respect to φ is equal to

$$2\int_0^\pi e^{-i\sigma\rho\cos\varphi}\,d\varphi = 2\pi J_0(\sigma\rho)$$

(e.g. see [107], Sec. XVII.24). We see that function (18) is also independent of the polar angle, i.e. $F = F(\sigma)$ where

$$F(\sigma) = \frac{1}{2\pi}\int_0^\infty J_0(\sigma\rho)\,f(\rho)\,\rho\,d\rho \tag{20}$$

Analogously, formula (19) yields

$$f(\rho) = 2\pi\int_0^\infty J_0(\sigma\rho)\,F(\sigma)\,\sigma\,d\sigma \tag{21}$$

Formulas (20) and (21) determine the **Hankel transformation** and its inverse. Usually these formulas are written without the factors $\frac{1}{2\pi}$ and 2π because the combination $2\pi F(\sigma)$ can again be denoted as $F(\sigma)$. The term "Hankel's transformation" is also applied to the transformation which is obtained if, for any fixed $n = 0, 1, 2, \ldots$, we put $f = f_1(\rho)\,e^{in\varphi}$. Let the reader show that in this case we have $F = F_1(\sigma)\,e^{in\psi}$ where $f_1(\rho)$ and $F_1(\sigma)$ are connected by the formulas

$$F_1(\sigma) = \frac{1}{2\pi i^n}\int_0^\infty J_n(\sigma\rho)\,f_1(\rho)\,\rho\,d\rho \quad\text{and}\quad f_1(\rho) = 2\pi i^n\int_0^\infty J_n(\sigma\rho)\,F_1(\sigma)\,\sigma\,d\sigma$$

The properties of Hankel's transformation and formulas for the transforms of concrete functions can either be derived directly or by resorting to Fourier's transformation (18), (19) because

$$f(\rho) = f\left(\sqrt{x^2+y^2}\right) \quad\text{and}\quad F(\sigma) = F\left(\sqrt{\xi^2+\eta^2}\right)$$

4. Integral Transformations on a Finite Interval. Integral transformations can also be defined for a finite interval of integration. For this purpose we usually take the same kernels as indicated above but with finite limits a and b in formula (12). On the pro-

perties and applications of such transformations we refer the reader to [129], Chapter VI. As an example, let us consider Fourier's sine transformation for the interval $[0, \pi]$:

$$F(p)=\int_0^{\pi} \sin pt \cdot f(t)\,dt \tag{22}$$

This relationship between the functions $f(t)$ and $F(p)$ we again designate as $f(t) \longrightarrow F(p)$. The upper limit of integration may be other than π; for instance, prove that if

$$\Phi(p)=\int_0^{l} \sin pt \cdot f(t)\,dt$$

then

$$\frac{l}{\pi} f\left(\frac{l}{\pi} t\right) \longrightarrow \Phi\left(\frac{\pi}{l} p\right)$$

If the function $f(t)$ in formula (22) is absolutely integrable on the interval $0 \leqslant t \leqslant \pi$, the function $F(p)$ is analytic in the whole p-plane and hence it is an entire function. The formula

$$f(t)=\sum_{k=1}^{\infty} b_k \sin kt$$

expressing the expansion of the function $f(t)$ into Fourier's sine series $\left(\text{where } b_k=\frac{2}{\pi}\int_0^{\pi} f(t) \sin kt\,dt\right)$ can be written in the form

$$f(t)=\frac{2}{\pi}\sum_{k=1}^{\infty} F(k) \sin kt \qquad (0 \leqslant t \leqslant \pi) \tag{23}$$

and interpreted as the inversion formula for transformation (22).

Let us denote the transform of the function $f''(t)$ by $F_2(p)$; then we have

$$F_2(p)=\int_0^{\pi} \sin pt \cdot f''(t)\,dt = \sin pt \cdot f'(t)\Big|_{t=0}^{\pi} - p\int_0^{\pi} \cos pt \cdot f'(t)\,dt =$$

$$= \sin \pi p \cdot f'(\pi) - p \cos pt \cdot f(t)\Big|_{t=0}^{\pi} - p^2 \int_0^{\pi} \sin pt \cdot f(t)\,dt =$$

$$= -p^2 F(p) + \sin \pi p \cdot f'(\pi) - p \cos \pi p \cdot f(\pi) + pf(0) \tag{24}$$

where $f'(\pi)$ is understood as $f'(\pi-0)$ and the like.

Other properties of transformation (22) can either be proved directly or by taking into account that the function $F(p)$ in for-

mula (22) is the ordinary Fourier's sine transform of the function

$$\tilde{f}(t)=\begin{cases} f(t) & \text{for } 0\leqslant t\leqslant \pi \\ 0 & \text{for } \pi\leqslant t<\infty \end{cases}$$

This approach is also applicable to other integral transformations on a finite interval.

As in § 2, the properties of transformations on a finite interval can be used for solving various equations. For instance, let it be necessary to solve equation (2.14) for $0\leqslant t<\infty$, $0\leqslant x\leqslant \pi$ if initial condition (2.15) (for the interval $0\leqslant x\leqslant \pi$) and the boundary conditions

$$u\big|_{x=0}=\psi(t),\quad u\big|_{x=l}=\chi(t)\qquad (0\leqslant t<\infty)$$

are given (i.e. the functions $\varphi(x)$, $\psi(t)$ and $\chi(t)$ are known). Applying to $u(t, x)$ a transformation of type (22) with respect to x (which we designate as $u(t, x)\longrightarrow U(t, p)$) and considering t to be a parameter we obtain, by formulas (24) and (2.14), the equation

$$\frac{\partial U}{\partial t}=-ap^2U+\sin \pi p\cdot \frac{\partial u}{\partial x}\Big|_{x=\pi}-p\cos \pi p\cdot \chi(t)+p\psi(t) \qquad (25)$$

while the initial condition takes the form

$$U\big|_{t=0}=\Phi(p)=\int_0^{\pi}\sin px\cdot\varphi(x)\,dx \qquad (26)$$

In order to apply inversion formula (23) let us put $p=k$ in (25) where k is an arbitrary integer ($k=1, 2, 3, \ldots$). Then the second term on the right-hand side of (25) vanishes and we arrive at an ordinary first-order differential equation whose solution, for initial condition (26), is

$$U=U(t, k)=\Phi(k)\,e^{-ak^2t}+k\int_0^t[(-1)^{k+1}\chi(\tau)+\psi(\tau)]e^{-ak^2(t-\tau)}d\tau$$

(check it up!). The sought-for solution $u(t, x)$ is now obtained by an inversion formula of type (23):

$$u(t, x)=\frac{2}{\pi}\sum_{k=1}^{\infty}U(t, k)\sin kx$$

CHAPTER IV

Linear Algebra

The ideas and methods of linear algebra are widely spread in modern applied mathematics. Linear algebra is a powerful tool for a standard approach to various linear problems in physics and mathematics. Many nonlinear problems can also be treated by means of methods of linear algebra with the aid of linearization; some nonlinear methods can be successfully tested on linear models. Therefore, an engineer or an applied scientist must have a good command of basic notions and techniques of linear algebra. The necessary preliminaries for studying the subject matter of the present chapter are fundamental concepts of linear algebra such as determinants, systems of linear algebraic equations, matrices and quadratic forms, linear spaces and linear mappings (e.g. see [107], §§ VI.1, 2, VII.1-3, 6 and XI.1-3).

Here we shall more thoroughly study some divisions of the theory concerning linear mappings, quadratic forms and other objects of linear algebra which are important for applications. For more detail we refer the reader to [39], [47] and [118].

We remind the reader that a *linear space* is defined as a set (R) of some objects on which the *linear operations* (i.e. the addition of the objects and their multiplication by numbers) can be performed in such a way that certain requirements known as the *axioms of linear space* are satisfied. Depending on whether these objects (the *elements* or *vectors* of the space (R)) are multiplied only by real numbers or also by complex ones we speak about a *linear space over the coefficient field of real or complex numbers* or, shortly, about a *real or complex linear space*. We shall only deal with finite-dimensional spaces. In an n-dimensional linear space (R) any set of n linearly independent vectors is called a *basis*, and every element $\boldsymbol{x} \in (R)$ (remember that the symbol "$\in$" is read "belongs to") can be resolved in a unique manner with respect to the base vectors.

A linear space is said to be *Euclidean* if the notion of *scalar product* of any two elements $\boldsymbol{x}$ and $\boldsymbol{y}$ (which will be denoted as

$\boldsymbol{x}\cdot\boldsymbol{y}$ or $(\boldsymbol{x}, \boldsymbol{y})$) satisfying certain axioms is introduced for the elements. A Euclidean space can also be *real* or *complex*. The set of all *n-dimensional number vectors*, that is of column matrices (vectors) composed of n numbers, is a standard example of an n-dimensional Euclidean space in which the scalar product is determined by the rule

$$\text{if } \mathbf{x}=\begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} \text{ and } \mathbf{y}=\begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \text{ then}$$

$$(\mathbf{x}, \mathbf{y})=\begin{cases} x_1y_1+\ldots+x_ny_n \text{ for real space} \\ x_1y_1^*+\ldots+x_ny_n^* \text{ for complex space} \end{cases}$$

where the asterisk designates the conjugate complex numbers. (In this chapter we shall give number vectors in medium-faced type in contrast to elements of an arbitrary linear space which will be given in medium-faced italic print.) Both for real and complex Euclidean spaces we have the formula $\mathbf{x}\cdot\mathbf{y}=(\mathbf{x}, \mathbf{y})=\mathbf{y}^*\mathbf{x}$ where the asterisk indicates the *conjugate matrix* which in the real case is obtained from the given matrix by passing to its *transpose* and in the complex case by the additional replacement of all the elements of the transpose by their complex conjugates. All the other n-dimensional linear or Euclidean spaces are equivalent to the n-space of number vectors in the sense that they are *isomorphic to it*. For a Euclidean space we can introduce, in a natural manner, the notions of *orthogonal* (mutually perpendicular) vectors, of the *norm (length)* of a vector and of *Euclidean bases* (also termed *Cartesian*) consisting of mutually orthogonal unit vectors (*orthonormal* vectors).

One of the most important notions of linear algebra is that of a *linear mapping* $\boldsymbol{y}=\mathrm{A}\boldsymbol{x}$ $(\boldsymbol{x}\in(R),\ \boldsymbol{y}\in(S))$ *of a linear space* (R) *into a linear space* (S). It is meant that $\boldsymbol{x}$ runs over the whole space (R) and that the correspondence determined by the mapping $\boldsymbol{y}=\mathrm{A}\boldsymbol{x}$ satisfies the conditions $\mathrm{A}(\boldsymbol{x}_1+\boldsymbol{x}_2)=\mathrm{A}\boldsymbol{x}_1+\mathrm{A}\boldsymbol{x}_2$ and $\mathrm{A}(\lambda\boldsymbol{x})=\lambda\mathrm{A}\boldsymbol{x}$. The set $\mathrm{A}(R)$ of all images obtained when $\boldsymbol{x}$ runs through (R) does not necessarily coincide with (S); in the general case it can be a *(linear) subspace* of (S). If $\mathrm{A}(R)=S$ we speak about "a mapping of (R) *onto* (S)" and if $\mathrm{A}(R)$ is a (proper) subspace of (S) about "a mapping of (R) *into* (S)".

If there is a mapping $\boldsymbol{y}=\mathrm{A}\boldsymbol{x}$ of (R) into or onto (S), and in each space (R) and (S) some bases are chosen, every vector $\boldsymbol{x}$ is completely characterized by the corresponding number vector $\mathbf{x}$ composed of its coordinates with respect to the basis taken in (R) and, similarly, every vector $\boldsymbol{y}$ is uniquely determined by the

corresponding number vector y. Then we have the relation $\mathbf{y}=\mathbf{Ax}$ where A is the *matrix of the mapping* A *relative to the chosen bases.*

§ 1. CONJUGATE MAPPINGS

1. Direct Sum. Let (R_1), (R_2), ..., (R_m) be some subspaces of an n-dimensional linear space (R), the dimension of (R_k) $(k=1, 2, \ldots, m)$ being $n_k \geqslant 0$. Then if every vector $\boldsymbol{x} \in (R)$ can be represented in a unique manner in the form

$$\boldsymbol{x}=\boldsymbol{x}_1+\boldsymbol{x}_2+\ldots+\boldsymbol{x}_m \text{ where } \boldsymbol{x}_1 \in (R_1),\ \boldsymbol{x}_2 \in (R_2),\ \ldots,\ \boldsymbol{x}_m \in (R_m) \tag{1}$$

we say that (R) is the **direct sum** of the linear subspaces (R_k), $k=1, 2, \ldots m$ (the **direct decomposition** into the subspaces (R_k)).

For instance, in the three-dimensional geometric space we use the resolution of a vector along three axes (which corresponds to the case $m=3$, $n_1=n_2=n_3=1$) and also the resolution into components along an axis and a plane non-parallel to the axis (that is $m=2$, $n_1=1$, $n_2=2$).

It can easily be shown that direct decomposition (1) possesses the following properties:

1. *Any system* $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_m$ *of nonzero vectors* $\boldsymbol{a}_k \in (R_k)$ $(k=1, 2, \ldots, m)$ *is linearly independent.* In fact, if, for instance, we suppose that $\boldsymbol{a}_1=\alpha\boldsymbol{a}_2+\ldots+\gamma\boldsymbol{a}_m$ then, putting $\boldsymbol{x}_1=\boldsymbol{a}_1$, $\boldsymbol{x}_2=-\alpha\boldsymbol{a}_2, \ldots, \boldsymbol{x}_m=-\gamma\boldsymbol{a}_m$ we obtain the relation

$$\boldsymbol{x}_1+\boldsymbol{x}_2+\ldots+\boldsymbol{x}_m=\boldsymbol{0}=\boldsymbol{0}+\boldsymbol{0}+\ldots+\boldsymbol{0}$$

which contradicts the fact that the vector $\boldsymbol{0}$ is uniquely representable in form (1) (because $\boldsymbol{x}_1 \neq \boldsymbol{0}$).

2. *The dimensions* n_k *of the subspaces* (R_k) $(k=1, 2, \ldots, m)$ *satisfy the relation* $n_1+n_2+\ldots+n_m=n$. Indeed, if we choose a basis consisting of n_k vectors in every space (R_k) $(k=1, 2, \ldots, m)$, the collection of all these base vectors is a basis in (R) (why?), and therefore their total number is equal to n.

Conversely (we leave the proof to the reader), if we are given some subspaces (R_1), (R_2), ..., (R_m) of (R) and it is known that the first of the above properties takes place, the set $(\overline{R})$ of all vectors $\boldsymbol{x}$ of form (1) is a linear subspace (of the space (R)) representable as the direct sum of its subspaces (R_1), (R_2), ..., (R_m). If, in addition, the second property also holds, we have $(\overline{R})=(R)$.

2. Invariant Subspaces. Let A be a mapping of a space (R) into itself. A subspace (R_1) of the space (R) is said to be **invariant** with respect to the mapping A if under this mapping (R_1) goes into itself, that is if for every $\boldsymbol{x} \in (R_1)$ we have $\mathrm{A}\boldsymbol{x} \in (R_1)$. In what follows we shall exclude the trivial case when the space (R_1) is zero-dimensional, i.e. consists of a single zero vector.

As an example, consider a rotation of the three-dimensional geometric space about an axis. The set of all vectors parallel to the axis of rotation is a one-dimensional subspace invariant under the rotation. All the vectors perpendicular to the axis form a two-dimensional invariant subspace, and the collection of all the vectors of the space is a three-dimensional invariant subspace because every space can be regarded as its own subspace. Now suppose that there is a mapping A of a linear space into itself and consider the set of all eigenvectors corresponding to a given eigenvalue of A. This set is obviously an invariant subspace under the mapping A (why?).

Assume now that a space (R) is the direct sum of its subspaces (R_k) of dimensions n_k $(k=1, 2, \ldots, m)$ each of which is invariant with respect to a mapping A of (R) into itself. Let us take, in (R), a basis $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_n$ such that its first n_1 vectors belong to (R_1), the subsequent n_2 vectors belong to (R_2), etc. The matrix $\mathbf{A}$ of the mapping A relative to that basis has a special structure. Namely, as is known, the jth column of the matrix $\mathbf{A}$ is a column vector formed of the coordinates (in the chosen basis) of the vector $\mathrm{A}\boldsymbol{l}_j$. Consequently, in the first n_1 columns only the elements with numbers $1, 2, \ldots, n_1$ (counted from the top) can be nonzero (why?), in the subsequent n_2 columns only the elements with numbers $n_1+1, n_1+2, \ldots, n_1+n_2$ can be distinct from zero, etc. We thus see that the matrix $\mathbf{A}$ in this basis *is partitioned into blocks* so that it can be (conditionally) written in the form

$$\mathbf{A}=\begin{pmatrix} \mathbf{A}_1 & & & 0 \\ & \mathbf{A}_2 & & \\ & & \ddots & \\ 0 & & & \mathbf{A}_m \end{pmatrix} \tag{2}$$

with square **submatrices (blocks)** $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m$ along the principal diagonal and all other elements equal to zero. A matrix of form (2) is called **quasi-diagonal** and can be characterized by the set of numbers $(n_1, n_2, \ldots, n_m)$ where the bracketed integers are the orders of the submatrices $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_m$. For instance, the quasi-diagonal matrices

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

are characterized, respectively, by the sets (1, 1, 1), (1, 2) and (3) (by the way, the first matrix can also be regarded as a quasi-diagonal matrix with the orders (1, 2), (2, 1) or (3) of the constituent submatrices).

Conversely, if the matrix of a mapping A of a space (R) into itself is of form (2) relative to a certain basis, the subspace (R_1) **spanned** by the first n_1 base vectors (that is generated by these vectors in the sense that it consists of all their linear combinations), the space (R_2) spanned by the subsequent n_2 base vectors, etc. are all invariant subspaces under the mapping A.

From (1) we obtain

$$\mathrm{A}\boldsymbol{x} = \mathrm{A}\boldsymbol{x}_1 + \mathrm{A}\boldsymbol{x}_2 + \ldots + \mathrm{A}\boldsymbol{x}_m$$

and therefore, if the action of the mapping A in each subspace (R_k) is known, that is if the structures of the corresponding submatrices are given, we can judge upon the structure of the whole mapping A because it reduces to independent operations on the components of the vectors within the invariant **direct summands** (R_1), (R_2), ..., (R_m).

3. Conjugate Mappings. Take two complex Euclidean spaces (R) and (S) (for real spaces all the facts given below are considered quite similarly), and let A be a linear mapping of the space (R) into (S). In this case we shall briefly write $(R) \xrightarrow{\mathrm{A}} (S)$ or $\mathrm{A}:(R) \to (S)$ (but the reader must not confuse these expressions with formulas involving a passage to a limit!). Suppose that there exists a linear mapping $\mathrm{B}:(S) \to (R)$ which satisfies the condition

$$(\mathrm{A}\boldsymbol{x}, \boldsymbol{y})_{(S)} = (\boldsymbol{x}, \mathrm{B}\boldsymbol{y})_{(R)} \text{ for all } \boldsymbol{x} \in (R), \ \boldsymbol{y} \in (S) \qquad (3)$$

where the subscripts (R) and (S) indicate the spaces in which the corresponding scalar products are taken. The mapping B is then called the **conjugate (dual) mapping of** A and is denoted by the symbol A^*. We stress that if A *is a mapping of* (R) *into* (S) *then* A^* *maps* (S) *into* (R).

Let us take a Euclidean basis $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$ in (R) and a Euclidean basis $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_m$ in (S). Then the mapping A has a matrix **A** relative to these bases and the mapping B a matrix **B**. It can easily be proved that *for relation* (3) *to hold it is necessary and sufficient that* $\mathbf{B} = \mathbf{A}^*$; in other words, *the conjugate mappings have conjugate matrices in the Euclidean bases and vice versa.* Indeed, according to the rule of constructing the matrix of a mapping for given bases (e.g. [107], Sec. XI.6), the elements of the matrix **A** are the numbers $a_{jk} = (\mathrm{A}\boldsymbol{p}_k, \boldsymbol{q}_j)$ $(j = 1, 2, \ldots, n;\ k = 1, 2, \ldots, m)$ and the elements of the matrix **B** are the numbers $(\mathrm{B}\boldsymbol{q}_k, \boldsymbol{p}_j)$ $(j = 1, 2, \ldots, n;\ k = 1, 2, \ldots, m)$. Hence, by (3), we have $b_{jk} = \mathrm{B}\boldsymbol{q}_k, \boldsymbol{p}_j) = (\mathrm{A}\boldsymbol{p}_j, \boldsymbol{q}_k)^* = a_{kj}^*$, i.e. $\mathbf{B} = \mathbf{A}^*$. Conversely, if $\mathbf{B} = \mathbf{A}^*$, $\boldsymbol{x} = \sum_{r=1}^{n} \alpha_r \boldsymbol{p}_r$, and $\boldsymbol{y} = \sum_{s=1}^{m} \beta_s \boldsymbol{q}_s$, the left member of (3) is equal to $\sum_{r,s} \alpha_r \beta_s^* a_{sr}$ and the right member is equal to $\sum_{r,s} \alpha_r \beta_s^* b_{rs}^* = \sum_{r,s} \alpha_r \beta_s^* a_{sr}$.

The above proof indicates that for every mapping there exists its conjugate mapping which is unique (namely, the mapping with the conjugate matrix in every Euclidean basis); the mapping conjugate to the conjugate mapping of A coincides with the original mapping A, that is $A^{**}=(A^*)^*=A$, and hence the mappings A^* and A are mutually conjugate.

If $(S)=(R)$, i.e. A is a linear mapping (operator) from (R) into itself, the conjugate mapping A^* is termed the **adjoint mapping** of A. If $(S)=(R)$ and $A^*=A$ the mapping A is said to be **self-adjoint**. In every Euclidean basis the matrix **A** of such a mapping satisfies the equality $\mathbf{A}=\mathbf{A}^*$, i.e. $a_{jk}=\overline{a}_{kj}$ $(j, k=1, 2, \ldots, n)$, and is called **self-adjoint** or **Hermitian matrix** (after C. Hermite, 1822-1901, a French mathematician). For real matrices the notion of a self-adjoint matrix is equivalent to that of a **symmetric matrix** (coinciding with its transpose) for which $a_{jk}=a_{kj}$.

Geometric examples illustrating the properties of conjugate mappings of real spaces can easily be obtained by taking the transposes of the corresponding matrices. For instance, let us consider mappings of a plane into itself. Then it is readily seen that a stretching (or shrinking) along an axis, a uniform magnification, a reflection through an axis and a projection on a straight line are examples of self-adjoint mappings. But a rotation through an arbitrary angle α and a shift along an axis are not self-adjoint: the conjugate (adjoint) mapping of the former is the rotation through the angle $-\alpha$ and that of the latter is a similar shift along the perpendicular axis.

4. Decomposition Connected with Conjugate Mappings. A direct decomposition (see Sec. 1) of a Euclidean space is said to be **orthogonal** if any two vectors taken from two different direct summands are orthogonal. For example, the three-dimensional geometric space can be represented as the orthogonal direct sum of a plane and a straight line perpendicular to it.

For any subspace (R_1) of a Euclidean space (R) there exists the so-called **orthogonal complement** which is a subspace (R_2) of (R) such that (R) is the orthogonal direct sum of (R_1) and (R_2), and (R_2) is uniquely determined by (R_1). To construct (R_2) for a given (R_1) it is sufficient to take the set of all vectors of (R) each of which is orthogonal to the whole subspace (R_1), that is to all the vectors belonging to (R_1) (the uniqueness of (R_2) is obvious; let the reader prove it). This construction also shows that if (R_2) is the orthogonal complement of (R_1), then (R_1) is the orthogonal complement of (R_2); in particular, $\boldsymbol{0}$ and (R) are the orthogonal complements of each other.

Let there be given a linear mapping A of a space (R) into a space (S). The set of all vectors $\boldsymbol{x}\in(R)$ for which $A\boldsymbol{x}=\boldsymbol{0}$ is called the **null space of the mapping** A and is denoted as $A^{-1}(\boldsymbol{0})$. The

set of all vectors of the form $A\boldsymbol{x}$ ($\boldsymbol{x} \in (R)$) is said to be the **image** of (R) under the mapping A; it is denoted by the symbol $A(R)$ we have already used. We suggest that the reader prove that $A^{-1}(\boldsymbol{0})$ is a linear subspace of (R) and $A(R)$ is a linear subspace of (S), the dimension of $A(R)$ being equal to the rank of the matrix **A** of the mapping A irrespective of the choice of the bases in the spaces (R) and (S).

Let us now suppose that the spaces (R) and (S) are Euclidean. *Then (R) is the orthogonal direct sum of its subspaces* $A^{-1}(\boldsymbol{0})$ *and* $A^*(S)$.

Indeed, we can readily prove that these subspaces are orthogonal to each other: if $\boldsymbol{x}_1 \in A^{-1}(\boldsymbol{0})$ and $\boldsymbol{x}_2 \in A^*(S)$ then, by definition, we have $A\boldsymbol{x}_1 = \boldsymbol{0}$ and $\boldsymbol{x}_2 = A^*\boldsymbol{y}$ ($\boldsymbol{y} \in (S)$), and therefore formula (3) implies that $(\boldsymbol{x}_1, \boldsymbol{x}_2)_{(R)} = (\boldsymbol{x}_1, A^*\boldsymbol{y})_{(R)} = (A\boldsymbol{x}_1, \boldsymbol{y})_{(S)} = (\boldsymbol{0}, \boldsymbol{y})_{(S)} = 0$. Now it only remains to show that every vector $\boldsymbol{x} \in (R)$ is representable in the form of a sum $\boldsymbol{x}_1 + \boldsymbol{x}_2$ where $\boldsymbol{x}_1 \in A^{-1}(\boldsymbol{0})$ and $\boldsymbol{x}_2 \in A^*(S)$. To this end, let us denote the projection of the vector $\boldsymbol{x}$ on the subspace $A^*(S)$ by $\boldsymbol{x}_2$; then we have $(\boldsymbol{x} - \boldsymbol{x}_2, \boldsymbol{x}')_{(R)} = 0$ for any vector $\boldsymbol{x}' \in A^*(S)$. In other words, the relation $(\boldsymbol{x} - \boldsymbol{x}_2, A^*\boldsymbol{y})_{(R)} = 0$ holds for every $\boldsymbol{y} \in (S)$. On the basis of (3) we now conclude that $(A(\boldsymbol{x} - \boldsymbol{x}_2), \boldsymbol{y})_{(S)} = 0$ and hence, $A(\boldsymbol{x} - \boldsymbol{x}_2) = \boldsymbol{0}$ because $\boldsymbol{y}$ is taken quite arbitrarily. Consequently, we obtain the relation $\boldsymbol{x} - \boldsymbol{x}_2 = \boldsymbol{x}_1 \in A^{-1}(\boldsymbol{0})$, and thus the validity of the desired orthogonal direct decomposition has been proved.

The mappings A and A^* being mutually conjugate, the application of the above proof to A^* shows that the space (S) is also the orthogonal direct sum of its subspaces $A^{*-1}(\boldsymbol{0})$ and $A(R)$. In other words, *the equation* $A\boldsymbol{x} = \boldsymbol{b}$ *considered for a given* $\boldsymbol{b} \in (S)$ *possesses at least one solution* $\boldsymbol{x} \in (R)$ *if and only if* $\boldsymbol{b}$ *is orthogonal to* $A^{*-1}(\boldsymbol{0})$, *that is orthogonal to all linearly independent solutions of the equation* $A^*\boldsymbol{y} = \boldsymbol{0}$ ($\boldsymbol{y} \in (S)$). It turns out that this assertion also applies to a wide class of linear equations in infinite-dimensional spaces.

The orthogonal direct decomposition whose validity has been proved above implies, in particular, that

$$\dim A^{-1}(\boldsymbol{0}) + \dim A^*(S) = \dim (R), \quad \dim A^{*-1}(\boldsymbol{0}) + \dim A(R) = \dim (S)$$

where "dim" means "the dimension of". Besides, if we introduce in (R) and (S) Euclidean bases it becomes clear that we always have $\dim A(R) = \dim A^*(S)$ (because two mutually conjugate matrices are of the same rank). It follows that $\dim A^{-1}(\boldsymbol{0}) - \dim A^{*-1}(\boldsymbol{0}) = \dim (R) - \dim (S)$. In particular, if $(R) = (S)$, that is if (R) is mapped into itself, we have $\dim A^{-1}(\boldsymbol{0}) = \dim A^{*-1}(\boldsymbol{0})$, and consequently *the maximum number of linearly independent solutions of the equation* $A\boldsymbol{x} = \boldsymbol{0}$ *is equal to that of the equation* $A^*\boldsymbol{y} = \boldsymbol{0}$.

5. Mapping of a Space into Itself. If $(R) = (S)$ we can speak about the *eigenvectors* and *eigenvalues* of a mapping A of (R) into

itself and also about those of the mapping A*. It is readily proved that *if* λ *is an eigenvalue of the mapping* A, *the number* λ^* *is an eigenvalue of the mapping* A* *of the same multiplicity.* In fact, choosing in (R) a Euclidean basis we see that the corresponding *characteristic equations* $\det(\mathbf{A}-\lambda\mathbf{I})=0$ and $\det(\mathbf{A}^*-\lambda\mathbf{I})=0$ go into one another if all the coefficients are replaced by their complex conjugates, and therefore the assertion under consideration is implied by the properties of conjugate complex numbers (e.g. see [107], Secs. VIII.3 and 8).

We shall now establish the following orthogonality property: *if* $\boldsymbol{p}$ *and* λ *are, respectively, an eigenvector and the corresponding eigenvalue of the mapping* A, $\boldsymbol{q}$ *and* μ *are the same for the mapping* A* *and* $\lambda \neq \mu^*$, *the vectors* $\boldsymbol{p}$ *and* $\boldsymbol{q}$ *are orthogonal*, i.e. $(\boldsymbol{p}, \boldsymbol{q})=0$. This assertion follows immediately from the comparison of the rightmost and leftmost members in the equalities

$$\lambda(\boldsymbol{p}, \boldsymbol{q})=(\lambda\boldsymbol{p}, \boldsymbol{q})=(\mathrm{A}\boldsymbol{p}, \boldsymbol{q})=(\boldsymbol{p}, \mathrm{A}^*\boldsymbol{q})=(\boldsymbol{p}, \mu\boldsymbol{q})=\mu^*(\boldsymbol{p}, \boldsymbol{q})$$

The above property implies the following corollary. Let all the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of the mapping A be simple (not multiple, i.e. distinct from each other) and $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$ be the corresponding nonzero eigenvectors; then, as is known (e.g. see [107], Sec. XI.4), they must be linearly independent and therefore can be taken as a basis in (R). We showed above that $\lambda_1^*, \lambda_2^*, \ldots, \lambda_r^*$ are eigenvalues of the mapping A*. Let $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n$ be the corresponding nonzero eigenvectors (which also form a basis in (R)). Then these two bases of the space (R) satisfy the condition

$$(\boldsymbol{p}_j, \boldsymbol{q}_k)=0 \qquad (j\neq k;\ j, k=1, \ldots, n) \tag{4}$$

Two bases satisfying condition (3) are termed **biorthogonal** and thus the bases $\boldsymbol{p}_j$ and $\boldsymbol{q}_j$ are biorthogonal.

If we consider the expansion of a vector with respect to a given basis the coefficients of the expansion can be easily determined if the corresponding biorthogonal basis is known. In fact, multiplying scalarly the equality

$$\boldsymbol{x}=\alpha_1\boldsymbol{p}_1+\alpha_2\boldsymbol{p}_2+\ldots+\alpha_n\boldsymbol{p}_n$$

by $\boldsymbol{q}_k$ and taking advantage of formula (4) we obtain

$$(\boldsymbol{x}, \boldsymbol{q}_k)=\alpha_k(\boldsymbol{p}_k, \boldsymbol{q}_k), \text{ that is } \alpha_k=\frac{(\boldsymbol{x}, \boldsymbol{q}_k)}{(\boldsymbol{p}_k, \boldsymbol{q}_k)} \quad (k=1, 2, \ldots, n)$$

It is clear that an orthogonal basis is one that is biorthogonal to itself.

Given a basis $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n$ (which does not necessarily consist of eigenvectors of a mapping), it is possible to construct its biorthogonal basis $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n$ which is determined uniquely

to within nonzero scalar factors and the order of the vectors. To this end, we can take as q_1 a vector which is orthogonal to the $(n-1)$-dimensional subspace spanned by the vectors $p_2, \ldots, p_n$, etc. Then it can be shown (let the reader do it) that the vectors $q_1, q_2, \ldots, q_n$ thus constructed are linearly independent and hence form a basis satisfying condition (4).

For an n-dimensional Euclidean space it is possible to introduce the notion of a vector product $[x_1, x_2, \ldots, x_{n-1}]$ of $n-1$ vectors $x_1, \ldots, x_{n-1}$ whose properties are similar to those of the ordinary vector product of two vectors in the three-dimensional geometric space. Then, when constructing the biorthogonal basis to $p_1, \ldots, p_n$, we can simply put $q_1 = [p_2, p_3, \ldots, p_n]$, etc.

6. Self-Adjoint Mapping. Let A be a self-adjoint mapping of a space (R) into itself. This means that $A^* = A$ which is equivalent to the condition

$$(Ax, y) = (x, Ay) \text{ for all } x, y \in (R) \tag{5}$$

We remind the reader that for condition (5) to hold it is necessary and sufficient that the matrix **A** of the mapping A be Hermitian in every Euclidean basis, i.e. $\mathbf{A}^* = \mathbf{A}$.

We shall prove that *all the eigenvalues of a self-adjoint mapping are real.* Indeed, if we have $Ax_1 = \lambda x_1$ $(x_1 \neq 0)$, then substituting $x = y = x_1$ into (5) we obtain, after cancelling by $(x_1, x_1) = |x_1|^2 \neq 0$, the equality $\lambda = \lambda^*$ which implies the validity of this assertion.

Now let us show that *the orthogonal complement of a subspace invariant under a self-adjoint mapping* A *is also invariant.* In fact, let (R_1) be an invariant subspace of (R) and (R_2) be the orthogonal complement of (R_1). Take an arbitrary vector $x \in (R_2)$. We must prove that $Ax \in (R_2)$ which is equivalent to the assertion that Ax is orthogonal to every vector $y \in (R_1)$. But this follows from (5) because $Ay \in (R_1)$.

The above properties imply the following consequence: *it is possible to construct in* (R) *a Euclidean basis composed of eigenvectors of a given self-adjoint mapping* A. Indeed, let us begin with choosing an eigenvector l_1 of unit length. It follows that the $(n-1)$-dimensional subspace (S) of all vectors orthogonal to l_1 is invariant, and therefore we can consider A as a mapping defined in (S) (in other words, we can take the *restriction* of A to (S)). Since the mapping A thus considered is a self-adjoint mapping of (S) into itself we can take a unit eigenvector $l_2 \in (S)$ of this mapping and then pass to the $(n-2)$-dimensional subspace of all vectors belonging to (S) which are orthogonal to l_2, etc. Proceeding in this way we finally arrive at the desired basis.

This consequence can be reformulated by using a purely matrix terminology. As is known, when one Euclidean basis is replaced

by another the matrix $\mathbf{A}'$ of a mapping A relative to the new basis is expressed, in terms of the matrix $\mathbf{A}$ of the mapping A in the old basis, by the formula $\mathbf{A}' = \mathbf{H}^{-1}\mathbf{A}\mathbf{H}$ where the transformation matrix $\mathbf{H}$ from the old basis to the new one satisfies the relation

$$\mathbf{H}^*\mathbf{H} = \mathbf{H}\mathbf{H}^* = \mathbf{I}, \text{ that is } \mathbf{H}^* = \mathbf{H}^{-1}$$

Matrices $\mathbf{H}$ of that kind are called **unitary** (in the case of a real space they are termed **orthogonal**). Thus we conclude that *for every Hermitian matrix* $\mathbf{A}$ *there is a unitary matrix* $\mathbf{H}$ *such that* $\mathbf{H}^{-1}\mathbf{A}\mathbf{H}$ *is a diagonal matrix with eigenvalues of the matrix* $\mathbf{A}$ *on its principal diagonal.*

The eigenvalues of a self-adjoint mapping being real, we also conclude that *the last two consequences are also valid for self-adjoint mappings of a real Euclidean space into itself and for real symmetric matrices.*

7. Extremal Properties of Eigenvalues. Let A be a self-adjoint mapping of a real space (R) into itself. Associating with every vector $\boldsymbol{x}$ the number $(\mathrm{A}\boldsymbol{x}, \boldsymbol{x})$ we obtain a scalar function defined on (R). Let us consider the values of this function assumed on the **unit sphere** of the space, that is on the $(n-1)$-dimensional manifold (S) of vectors satisfying the relation $|\boldsymbol{x}| = 1$. Then, *the eigenvectors of the mapping* A *are those for which the function* $(\mathrm{A}\boldsymbol{x}, \boldsymbol{x})$ *takes its stationary values on* (S), *and these stationary values are equal to the corresponding eigenvalues.* In particular, *the greatest and the least eigenvalues are, respectively, equal to the maximum and minimum values of the function* $(\mathrm{A}\boldsymbol{x}, \boldsymbol{x})$ *on* (S) *while for every eigenvector corresponding to an intermediate eigenvalue this function has a minimax.*

To prove this proposition let us take in (R) a Euclidean basis consisting of unit eigenvectors $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_n$ of the mapping A. Then every vector $\boldsymbol{x} \in (R)$ is expanded, relative to this basis, in the form $\boldsymbol{x} = x_1\boldsymbol{l}_1 + x_2\boldsymbol{l}_2 + \ldots + x_n\boldsymbol{l}_n$, and the function $(\mathrm{A}\boldsymbol{x}, \boldsymbol{x})$ can be written as

$$(\mathrm{A}\boldsymbol{x}, \boldsymbol{x}) = \lambda_1 x_1^2 + \lambda_2 x_2^2 + \ldots + \lambda_n x_n^2 \equiv f(x_1, x_2, \ldots, x_n)$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of the mapping A. The equation of the sphere (S), in the coordinates $x_1, x_2, \ldots, x_n$, takes the form

$$g(x_1, x_2, \ldots, x_n) \equiv x_1^2 + x_2^2 + \ldots + x_n^2 = 1$$

Thus we have arrived at the ordinary problem of finding a *conditional extremum* of the function $f(x_1, \ldots, x_n)$ with the *constraint (subsidiary condition)* $g(x_1, \ldots, x_n) = 1$. To determine the (conditional) stationary values we can apply the *method of Lagrange's*

multipliers and thus derive the equations

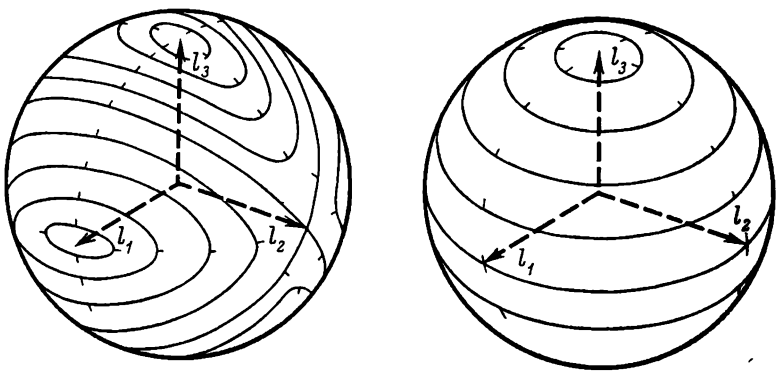
$$\frac{\partial}{\partial x_k}(f-\mu g) \equiv 2\lambda_k x_k - 2\mu x_k = 0 \qquad (k=1,\ 2,\ \ldots,\ n) \qquad (6)$$

where μ is the Lagrange multiplier. Since all x_k cannot turn into zero simultaneously, relations (6) imply that one of the differences $\lambda_k-\mu$, $k=1,\ 2,\ \ldots,\ n$, is equal to zero. Suppose that, for instance, $\lambda_k-\mu=0$ for $k=j$. Then it follows from (6) that all the coordinates x_k for which $\lambda_k \neq \lambda_j$ must be equal to zero, this condition specifying eigenvectors which correspond to the eigenvalue λ_j (why?).

The assertion concerning the maximum (and, similarly, the minimum) eigenvalue immediately follows if we consider all λ_k's as being arranged into the sequence $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$ and represent the function $(A\boldsymbol{x},\ \boldsymbol{x})$ in the form

$$(A\boldsymbol{x},\ \boldsymbol{x}) = \lambda_1 (x_1^2 + x_2^2 + \ldots + x_n^2) - (\lambda_1 - \lambda_2) x_2^2 - \\ - (\lambda_1 - \lambda_3) x_3^2 - \ldots - (\lambda_1 - \lambda_n) x_n^2$$

(let the reader carefully examine this argument). It should be noted that if the eigenvalue λ_1 is of multiplicity d where $d \geqslant 2$



(a) (b)

Fig. 63

the maximum of the function $(A\boldsymbol{x},\ \boldsymbol{x})$ attained for the vector $\boldsymbol{l}_1$ is nonstrict because this function takes on the constant value λ_1 on the whole $(d-1)$-dimensional intersection of (S) with the subspace spanned by the vectors $\boldsymbol{l}_1,\ \boldsymbol{l}_2,\ \ldots,\ \boldsymbol{l}_d$.

In Fig. 63 we see a possible disposition of level lines of the function $(A\boldsymbol{x},\ \boldsymbol{x})$ on the unit sphere of the three-dimensional geometric space on which the directions of decrease of the function are marked. Fig. 63*a* illustrates the case when $\lambda_1 > \lambda_2 > \lambda_3$

and Fig. 63*b* the case when $\lambda_1 = \lambda_2 > \lambda_3$ (let the reader carefully examine these figures).

The properties proved above are also valid for a self-adjoint mapping of a complex Euclidean space (R) into itself. The following assertion is sometimes of use for investigating these mappings: *a mapping* A *of such a space* (R) *into itself is self-adjoint if and only if all the values of the function* $(\mathrm{A}\boldsymbol{x}, \boldsymbol{x})$ *assumed in* (R) *are real.* Indeed, for every self-adjoint mapping A we have $(\mathrm{A}\boldsymbol{x}, \boldsymbol{x})^* = (\boldsymbol{x}, \mathrm{A}\boldsymbol{x}) = (\mathrm{A}\boldsymbol{x}, \boldsymbol{x})$, i.e. the values of $(\mathrm{A}\boldsymbol{x}, \boldsymbol{x})$ are real. If we take in (R) a basis formed of eigenvectors, the function $(\mathrm{A}\boldsymbol{x}, \boldsymbol{x})$ is expressed in the form $(\mathrm{A}\boldsymbol{x}, \boldsymbol{x}) = \lambda_1 |x_1|^2 + \lambda_2 |x_2|^2 + \ldots + \lambda_n |x_n|^2$, $(|x_1|^2 + |x_2|^2 + \ldots + |x_n|^2 = |\boldsymbol{x}|^2)$ (check it up!). This relation implies the validity of the above-mentioned properties for the case of a complex space. It can readily be shown that any mapping A can be represented in the form $\mathrm{A}_1 + i\mathrm{A}_2$ where $\mathrm{A}_1 = \frac{\mathrm{A} + \mathrm{A}^*}{2}$ and $\mathrm{A}_2 = \frac{\mathrm{A} - \mathrm{A}^*}{2i}$ are self-adjoint mappings. Therefore, conversely, if the values of the function $(\mathrm{A}\boldsymbol{x}, \boldsymbol{x}) = (\mathrm{A}_1\boldsymbol{x}, \boldsymbol{x}) + i(\mathrm{A}_2\boldsymbol{x}, \boldsymbol{x})$ are real we conclude that $(\mathrm{A}_2\boldsymbol{x}, \boldsymbol{x}) \equiv 0$ whence $\mathrm{A}_2\boldsymbol{x} \equiv \boldsymbol{0}$ (why?), and thus $\mathrm{A} = \mathrm{A}_1$ and $\mathrm{A} = \mathrm{A}^*$.

It is sometimes convenient to consider, instead of the function $(\mathrm{A}\boldsymbol{x}, \boldsymbol{x})$ $(\boldsymbol{x} \in (S))$, the normalized function $\varphi(\boldsymbol{x}) = \frac{(\mathrm{A}\boldsymbol{x}, \boldsymbol{x})}{(\boldsymbol{x}, \boldsymbol{x})}$ which is defined in the whole space (R) except the vector $\boldsymbol{x} = \boldsymbol{0}$ for which it has a discontinuity. The function φ being constant on every straight line passing through the origin, the above extremal properties (including stationarity) of the eigenvalues can also be formulated in terms of φ.

These extremal properties of the eigenvalues are used for estimation and approximate computation of the greatest and least eigenvalues and approximate computation of the corresponding eigenvectors of a given self-adjoint mapping or a given symmetric matrix by means of numerical solution of the corresponding extremal problem for a function of several variables. This method was inaugurated by J. Rayleigh (1842-1919), an English physicist.

To compute the eigenvalue λ_2 (we again suppose that the eigenvalues are taken as the sequence $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$) and the corresponding eigenvector $\boldsymbol{l}_2$ we can use the fact that, as is seen from the above considerations, the quantity λ_2 is equal to the maximum of the function $\varphi(\boldsymbol{x})$ assumed on the subspace of vectors satisfying the subsidiary condition $(\boldsymbol{x}, \boldsymbol{l}_1) = 0$. To compute λ_3 we make use of the two conditions $(\boldsymbol{x}, \boldsymbol{l}_1) = 0$ and $(\boldsymbol{x}, \boldsymbol{l}_2) = 0$, etc.

To evaluate and investigate the eigenvalues of a mapping it is sometimes advisable to apply the following proposition proved by R. Courant (and known as **Courant's minimum-maximum principle** or **minimax theorem**) which gives an expression for every eigen-

value irrelevant of the construction of the preceding eigenvalues:

$$\lambda_k = \min_{a_1, \ldots, a_{k-1} \in (R)} \left[\max_{\substack{(x, a_1) = \ldots = (x, a_{k-1}) = 0 \\ |x| = 1}} (Ax, x) \right] \quad (k = 1, \ldots, n) \tag{7}$$

In formula (7) it is supposed that the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are arranged in a monotone decreasing sequence, the maximum is taken, for chosen vectors $a_1, \ldots, a_{k-1}$, over all the unit vectors x orthogonal to them, and the resultant minimum is taken over all these maximum values when the collection of the vectors $a_1, \ldots, a_n$ runs through all the possible combinations of $k-1$ vectors belonging to the space (R).

To prove this theorem we note that, for given $a_1, \ldots, a_{k-1}$, among the vectors of the form $x = \alpha_1 l_1 + \ldots + \alpha_k l_k$ there is at least one satisfying the conditions $(x, a_1) = \ldots = (x, a_{k-1}) = 0$, $|x| = 1$ (why?). But, for this vector, we have $(Ax, x) = \lambda_1 \alpha_1^2 + \ldots + \lambda_k \alpha_k^2 \geqslant \lambda_k$, and therefore the whole right-hand side of (7) is not less than λ_k. Now we see that if we put $a_1 = l_1, \ldots, a_{k-1} = l_{k-1}$ the expression in the square brackets in (7) is equal to λ_k (check it up!), and hence the right-hand side of (7) is itself equal to λ_k.

§ 2. QUADRATIC FORMS

1. Introduction. In algebra a **form** of several variables is understood as a homogeneous polynomial expression in two or more variables. According to the degree of the polynomial a form is said to be **linear, quadratic,** etc. Here we shall dwell on quadratic forms in real variables with real coefficients. As is known, a quadratic form F of n variables $x_1, x_2, \ldots, x_n$ can be written in the form $F = \mathbf{x}^* \mathbf{A} \mathbf{x}$ where $\mathbf{x}$ is a column vector (matrix) composed of these variables and $\mathbf{A}$ is a symmetric matrix corresponding to the given form. Under a linear transformation of the variables performed according to the formula $\mathbf{x} = \mathbf{H}\mathbf{x}'$ (where $\mathbf{x}'$ is the column of the new variables $x_1', \ldots, x_n'$) the quadratic form is transformed by the formula $F = \mathbf{x}'^* \mathbf{A}' \mathbf{x}'$ where $\mathbf{A}' = \mathbf{H}^* \mathbf{A} \mathbf{H}$ (e.g. see [107], Sec. XI.11).

The results of Sec. 1.6 indicate that it is always possible to choose an orthogonal matrix $\mathbf{H}$ such that the matrix $\mathbf{A}'$ takes the diagonal form, that is the quadratic form is expressed in the new variables in the form

$$F = \lambda_1 x_1'^2 + \lambda_2 x_2'^2 + \ldots + \lambda_n x_n'^2 \tag{1}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of the original matrix $\mathbf{A}$. A quadratic form of type (5) not containing pairwise products of different variables is said to be in **diagonal (canonical) form.**

If all the eigenvalues of the matrix **A** are positive the form F is called **positive definite**; if all λ_k's are negative it is called **negative definite.** The forms of both types are referred to as **definite.** A feature of these forms is that they are equal to zero only when $x_1 = x_2 = \ldots = x_n = 0$ (why?). If among the eigenvalues λ_k there are some equal to zero the form is termed **degenerate.** For a form F to be degenerate it is necessary and sufficient that $\det \mathbf{A} = 0$. Formula (1) shows that, after an appropriate change of variables has been performed, every degenerate form of type F becomes a function of less than n independent variables.

The above **reduction of a quadratic form to canonical form** (1) admits of a simple geometric interpretation. The equation $F(x_1, x_2, \ldots, x_n) = 1$ defines in the n-dimensional Cartesian space E_n of number vectors a manifold (M) of dimension $n-1$ which can naturally be interpreted as a surface of the second order. Since we have $F(-x_1, -x_2, \ldots, -x_n) \equiv F(x_1, x_2, \ldots, x_n)$, the origin O is the centre of symmetry of (M). The change of variables according to the formula $\mathbf{x} = \mathbf{Hx}'$ with an orthogonal matrix **H** is equivalent to a rotation of the coordinate axes about O (e.g. see [107], Sec. XI.9). Therefore, the reduction to diagonal form corresponds to a rotation of the axes such that after the rotation has been performed the equation of the surface (M) takes the form

$$\lambda_1 x_1'^2 + \lambda_2 x_2'^2 + \ldots + \lambda_n x_n'^2 = 1 \qquad (2)$$

which is referred to as a **canonical (standard)** equation of the surface (M) (compare with the canonical equations of the lines and surfaces of the second order considered in analytic geometry). Equation (2) shows that the axes $x_1', x_2', \ldots, x_n'$ serve as symmetry axes of the surface (M); they are referred to as its **principal axes.** Therefore a transformation of a quadratic form to diagonal form is also called a **reduction to principal axes** (a **principal axes transformation).**

If a form F is non-degenerate then, depending on the number of positive and negative numbers λ_k, there can appear $n+1$ types of surfaces. In particular, if all λ_k's are positive the surface (M) is called an *ellipsoid* with semiaxes $\frac{1}{\sqrt{\lambda_k}}$ (why?) and if all the numbers λ_k are negative we obtain an imaginary surface. All the other cases correspond to *hyperboloids* of different types. (Let the reader interpret the surface (M) when the form F is degenerate.)

In conclusion let us consider the so-called **Hermitian quadratic forms** $F = \mathbf{z}^* \mathbf{Az}$ with an Hermitian matrix **A** (see Sec. 1.3) and complex column vector **z**. We have $F^* = F$ (why?), and therefore a form of this type only assumes real values. After the corresponding unitary transformation $\mathbf{z} = \mathbf{Hz}'$ (Sec. 1.6) has been made

it is reduced to the diagonal form

$$F = \mathbf{z}'^{*} \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)\, \mathbf{z}' = \\ = \lambda_1 |z_1'|^2 + \lambda_2 |z_2'|^2 + \ldots + \lambda_n |z_n'|^2$$

2. Law of Inertia for Quadratic Forms. Let us again consider a real quadratic form F $\mathbf{x}^{*}\mathbf{A}\mathbf{x}$. It can be transformed to diagonal form in many different ways without imposing the condition that the transformation matrix $\mathbf{H}$ is orthogonal (e.g. see Sec. 3). Then the coefficients in the squares of the variables are not invariant and depend on the way the reduction is performed. But it turns out that *the number of positive, negative and zero coefficients is independent of the transformation to canonical form.* This fact is known as the **(Sylvester)** *) **inertia law for quadratic forms.** The number of positive coefficients is called the **index** of the form, the number of positive coefficients diminished by the number of negative ones is termed the **signature** and the total number of nonzero coefficients is the **rank** of the form. According to the law of inertia, these characteristics of the form are invariant with respect to any invertible change of variables.

To prove the inertia law let us suppose that a given form F is reduced to diagonal form by means of two different transformations $\mathbf{x} = \mathbf{H}'\mathbf{x}'$ and $\mathbf{x} = \mathbf{H}''\mathbf{x}''$. Then we can equate the results and write

$$\lambda_1' x_1'^2 + \lambda_2' x_2'^2 + \ldots + \lambda_n' x_n'^2 \equiv \lambda_1'' x_1''^2 + \lambda_2'' x_2''^2 + \ldots + \lambda_n'' x_n''^2 \qquad (3)$$

We can limit ourselves to proving that the numbers of positive coefficients on the left-hand and right-hand sides are equal because changing the signs we obtain as a consequence that the numbers of negative coefficients are equal and therefore the numbers of zero coefficients are also the same.

Let us suppose that in relation (3) the coefficients $\lambda_1', \lambda_2', \ldots, \lambda_{k'}'$ on the left-hand side and the coefficients $\lambda_1'', \lambda_2'', \ldots, \lambda_{k''}''$ on the right-hand side are positive. For definiteness, let $k'' > k'$. Now we put

$$x_1' = 0,\ x_2' = 0,\ \ldots,\ x_{k'}' = 0,\ x_{k''+1}'' = 0,\ x_{k''+2}'' = 0,\ \ldots,\ x_n'' = 0$$

If in these relations we express all the variables x_j' and x_j'' in terms of $x_1, x_2, \ldots, x_n$ we obtain relative to x_i's a system of homogeneous linear equations with n unknowns, the number of equations being equal to $k' + (n - k'') < n$. A system of this type always possesses a nonzero (nontrivial) solution (why?). Finding the values of all x_j' and x_j'' corresponding to this nonzero solution and substituting them into (3) we see that the left member is nonpositive while the right member is positive (why?). It follows that $k' = k''$, which is what we set out to prove.

*) J. J. Sylvester (1814-1897), an English mathematician.

3. Jacobi Method and Sylvester Theorem. Here we shall describe the method of K.G.J. Jacobi which makes it possible to transform a real quadratic form $F = \mathbf{x}^* \mathbf{A} \mathbf{x}$ to diagonal form without solving algebraic equations. It is supposed that the so-called **principal minors** of the matrix $\mathbf{A}$ (of the form F)

$$\Delta_1 = a_{11}, \quad \Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \quad \ldots, \quad \Delta_n = \begin{vmatrix} a_{11} a_{12} & \ldots & a_{1n} \\ a_{21} a_{22} & \ldots & a_{2n} \\ \cdot & \cdot & \cdot \\ a_{n1} a_{n2} & \ldots & a_{nn} \end{vmatrix}$$

are all different from zero.

The transformation is looked for in the triangular form

$$\left.\begin{aligned} x_1 &= x_1' + \alpha_{12} x_2' + \alpha_{13} x_3' + \ldots + \alpha_{1n} x_n', \\ x_2 &= \qquad x_2' + \alpha_{23} x_3' + \ldots + \alpha_{2n} x_n' \\ x_3 &= \qquad\qquad x_3' + \ldots + \alpha_{3n} x_n' \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ x_n &= \qquad\qquad\qquad\qquad x_n' \end{aligned}\right\} \tag{4}$$

and must lead to the identity

$$\sum_{i,\, j=1}^{n} a_{ij} x_i x_j \equiv p_1 x_1'^2 + p_2 x_2'^2 + \ldots + p_n x_n'^2 \tag{5}$$

The unknown coefficients α_{ij} and p_i are found by a step-by-step procedure. Let us first put $x_2' = x_3' = \ldots = x_n' = 0$ in formulas (4) and (5). This results in

$$x_1 = x_1', \quad a_{11} x_1^2 \equiv p_1 x_1'^2$$

whence $p_1 = a_{11} = \Delta_1$. Next we put $x_3' = x_4' = \ldots = x_n' = 0$ in formulas (4) and (5), which yields the relations

$$\left.\begin{aligned} x_1 &= x_1' + \alpha_{12} x_2', \quad x_2 = x_2' \\ a_{11} x_1^2 &+ 2a_{12} x_1 x_2 + a_{22} x_2^2 \equiv p_1 x_1'^2 + p_2 x_2'^2 \end{aligned}\right\} \tag{6}$$

Substituting the first two formulas (6) into the third and equating the coefficient in $x_1' x_2'$ to zero we arrive at the equation

$$2a_{11}\alpha_{12} + 2a_{12} = 0$$

from which we can determine α_{12} because, by the hypothesis, $a_{11} \neq 0$. Applying to relations (6) formula $\mathbf{A}' = \mathbf{H}^* \mathbf{A} \mathbf{H}$ and taking into account that in this case we have $\det \mathbf{H} = \det \mathbf{H}^* = 1$ we obtain (let the reader carry out the calculations) the equality

$$p_1 p_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \Delta_2 \quad \text{whence} \quad p_2 = \frac{\Delta_2}{\Delta_1}$$

Furthermore, putting $x_4' = x_5' = \ldots = x_n' = 0$ and equating the coefficients in $x_1'x_3'$ and $x_2'x_3'$ to zero we obtain a system of two linear equations from which we can find α_{13} and α_{23} since α_{12} has already been determined. It can be proved that the determinant of this system is equal to Δ_2 and, since we have, by the hypothesis, $\Delta_2 \neq 0$, the quantities α_{13} and α_{23} can in fact be found. From the formula of transformation of the matrix of a quadratic form we derive the relation

$$p_1p_2p_3 = \Delta_3 \text{ whence } p_3 = \frac{\Delta_3}{\Delta_2}$$

Proceeding in this manner we finally determine the required transformation which reduces the original form to the diagonal form

$$\Delta_1 x_1'^2 + \frac{\Delta_2}{\Delta_1} x_2'^2 + \frac{\Delta_3}{\Delta_2} x_3'^2 + \ldots + \frac{\Delta_n}{\Delta_{n-1}} x_n'^2 \tag{7}$$

Corollary. *If $\Delta_k \neq 0$ for all $k = 1, 2, \ldots, n$ the number s of negative coefficients in formula (7) is equal to the number of changes of the sign in the numerical sequence $1, \Delta_1, \Delta_2, \ldots, \Delta_n$, the number of positive coefficients is equal to $n-s$ and there are no zero coefficients.* This assertion is directly implied by formula (7) (showing that the above number of changes of the sign is in fact equal to s) because, by the hypothesis, there are no zero coefficients in (7) since the condition $\Delta_n \neq 0$ means that the form is nondegenerate and therefore the other $n-s$ coefficients are positive.

Next we prove the following

Sylvester Theorem. *For a quadratic form $\mathbf{x}^* \mathbf{A}\mathbf{x}$ to be positive definite it is necessary and sufficient that all the principal minors of the matrix $\mathbf{A}$ be positive.* Indeed, if all Δ_k's are nonzero the assertion of the theorem follows from the above paragraph. Now suppose that one of the principal minors, say Δ_k, is equal to zero. This means that the sum $\sum_{i,j=1}^{k} a_{ij}x_ix_j$, as quadratic form in the variables $x_1, x_2, \ldots, x_k$, is degenerate, and therefore there exists a nonzero (nontrivial) set of values $x_1 = b_1, x_2 = b_2, \ldots, x_k = b_k$ which turns this form into zero. But then the original form vanishes for a nonzero combination $x_1 = b_1, \ldots, x_k = b_k, x_{k+1} = 0, \ldots, x_n = 0$ of the values of the variables and therefore it cannot be a definite one. The theorem has thus been proved. It should be noted that it also remains true for Hermitian (complex) quadratic forms.

4. Simultaneous Reduction of Two Quadratic Forms to Canonical Form. *Let there be given two real quadratic forms $F = \mathbf{x}^* \mathbf{A}\mathbf{x}$ and $G = \mathbf{x}^* \mathbf{B}\mathbf{x}$, the former being positive definite and the latter arbitrary. Then there is a real invertible (nondegenerate) transformation $\mathbf{x} = \mathbf{H}\mathbf{x}'$ which simultaneously reduces the first form to the sum of squares and the second to canonical form.*

To prove the proposition, let us first perform a transformation $\mathbf{x}=\mathbf{K}\mathbf{y}$ reducing the form F to a diagonal form $p_1y_1^2+p_2y_2^2+\ldots+p_ny_n^2$. Then all p_k, $k=1, 2, \ldots, n$, are positive, and therefore we can additionally apply the transformation $p_jy_j^2=y_j'^2$ $(j=1, 2, \ldots, n)$,

$$\text{i.e. } y_1=\frac{1}{\sqrt{p_1}}y_1',\ y_2=\frac{1}{\sqrt{p_2}}y_2',\ \ldots,\ y_n=\frac{1}{\sqrt{p_n}}y_n'$$

which can be written in the form $\mathbf{y}=\mathbf{L}\mathbf{y}'$ (where $\mathbf{L}$ is the corresponding transformation matrix). This reduces F to the form

$$F=y_1'^2+y_2'^2+\ldots+y_n'^2=\mathbf{y}'^*\mathbf{y}'$$

The resultant transformation $x=\mathbf{K}\mathbf{L}\mathbf{y}'$ changes the second quadratic form into $G=\mathbf{y}'^*\mathbf{L}^*\mathbf{K}^*\mathbf{B}\mathbf{K}\mathbf{L}\mathbf{y}'=\mathbf{y}'^*\mathbf{B}'\mathbf{y}'$ where $\mathbf{B}'=\mathbf{L}^*\mathbf{K}^*\mathbf{B}\mathbf{K}\mathbf{L}$ $(\mathbf{B}'^*=\mathbf{B}')$. Now let us reduce the quadratic form $\mathbf{y}'^*\mathbf{B}'\mathbf{y}'$ to canonical form by means of an orthogonal transformation $\mathbf{y}'=\mathbf{H}'\mathbf{x}'$. This finally yields

$$\mathbf{x}=\mathbf{K}\mathbf{L}\mathbf{H}'\mathbf{x}'=\mathbf{H}\mathbf{x}' \text{ where } \mathbf{H}=\mathbf{K}\mathbf{L}\mathbf{H}'$$

According to the choice of $\mathbf{H}'$, under this transformation the form G becomes diagonal, while the form F goes into $\mathbf{x}'^*\mathbf{x}'=x_1'^2+x_2'^2+\ldots+x_n'^2$ (why?). The assertion has thus been proved.

The coefficients of the resultant canonical form G can easily be determined. In fact, the equalities $\mathbf{H}^*\mathbf{A}\mathbf{H}=\mathbf{I}$ and $\mathbf{H}^*\mathbf{B}\mathbf{H}=\operatorname{diag}(g_1, g_2, \ldots, g_n)$ imply that

$$(\det \mathbf{H})^2 \det(\lambda\mathbf{A}-\mathbf{B})=(\lambda-g_1)(\lambda-g_2)\ldots(\lambda-g_n)$$

(why?). Consequently, the diagonal coefficients g_j are the roots of the equation $\det(\lambda\mathbf{A}-\mathbf{B})=0$. But we have $\lambda\mathbf{A}-\mathbf{B}=-\mathbf{A}(\mathbf{A}^{-1}\mathbf{B}-\lambda\mathbf{I})$, and therefore these are the eigenvalues of the matrix $\mathbf{A}^{-1}\mathbf{B}$. We have incidentally shown that, under the hypotheses assumed, the matrix $\mathbf{A}^{-1}\mathbf{B}$ has all its eigenvalues real although in the general case it can be nonsymmetric.

The latter fact has an important application in the theory of vibrations. Suppose that we have an *autonomous system* (i.e. one whose properties remain time invariant) with a finite number n of degrees of freedom. Let $q_1, q_2, \ldots, q_n$ be generalized coordinates of the system and a function $U=U(q_1, q_2, \ldots, q_n)$ be its potential energy. Consider a state of the system which is close to its equilibrium position $q_1=0, q_2=0, \ldots, q_n=0$. Then it can be shown that, to within infinitesimals of order higher than the second, the kinetic and potential energies of the system can be written as quadratic forms

$$T=\frac{1}{2}\sum_{i,j=1}^{n}a_{ij}\dot{q}_i\dot{q}_j \quad \text{and} \quad U=\frac{1}{2}\sum_{i,j=1}^{n}b_{ij}q_iq_j\left(\dot{q}_i=\frac{dq_i}{dt}\right) \tag{8}$$

the former being positive definite. If we make a linear transformation $\mathbf{q} = \mathbf{Hq}'$ (with constant coefficients) of the generalized coordinates, we obtain $\dot{\mathbf{q}} = \mathbf{H}\dot{\mathbf{q}}'$, which implies that the form T is transformed as if the coordinates were not differentiated. Hence, using the result proved above, we see that it is possible to pass to new generalized coordinates $q_1', q_2', \ldots, q_n'$ in which the kinetic and potential energies take the forms

$$T = \frac{1}{2}(\dot{q}_1'^2 + \dot{q}_2'^2 + \ldots + \dot{q}_n'^2) \quad \text{and} \quad U = \frac{1}{2}(\mu_1 q_1'^2 + \mu_2 q_2'^2 + \ldots + \mu_n q_n'^2) \tag{9}$$

These variables $q_1', \ldots, q_n'$ are called **normal coordinates** of the system. Expression (9) makes it possible to prove that if $\mu_k > 0$ for all $k = 1, 2, \ldots, n$ (that is the form U in (8) is positive definite) the equilibrium position $q_1 = 0, q_2 = 0, \ldots, q_n = 0$ is stable and if there are some negative μ_k's it is unstable.

§ 3. STRUCTURE OF LINEAR MAPPING

Let us investigate the structure of a linear mapping A of an n-dimensional complex linear space (R) into itself. It is known (e. g. see [107], Secs. XI.4, 8) that if this mapping has n different eigenvalues then to each value λ_k there corresponds only one linearly independent eigenvector $\boldsymbol{l}_k$. In the basis composed of the eigenvectors the matrix of the mapping takes the diagonal form $\operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, that is the coordinates of each vector $\boldsymbol{x} \in (R)$ are transformed according to the formulas

$$y_1' = \lambda_1 x_1', \quad y_2' = \lambda_2 x_2', \quad \ldots, \quad y_n' = \lambda_n x_n'$$

where the primes designate the coordinates relative to this basis. Hence, the mapping A reduces to a combination of λ_1-fold stretching in the direction of $\boldsymbol{l}_1$, λ_2-fold stretching in the direction of $\boldsymbol{l}_2$, etc.

But in the general case when the characteristic equation for the eigenvalues has multiple roots the structure of the mapping A becomes more complicated. Our aim is now to describe this structure.

1. Mapping with a Single Eigenvector. We first of all note that according to the fundamental theorem of algebra every mapping A possesses at least one eigenvalue with which a d-dimensional $(d \geqslant 1)$ subspace spanned by the corresponding linearly independent eigenvectors is associated. In this section we shall suppose that the mapping in question has only one eigenvalue (i. e. all its eigenvalues are coincident) and that $d = 1$, i. e. there exists a single eigenvector (determined uniquely to within an arbitrary nonzero scalar factor). Let us begin with the case $\lambda = 0$ in which the eigenvectors are determined by the equality $A\boldsymbol{x} = \boldsymbol{0}$.

For our further aims we need the following lemma: *if a mapping* A *has only a zero eigenvalue, and, for a vector* $x \in R$ *and an integer* k, *the elements* x, Ax, A^2x, ..., A^kx *are linearly dependent, then* $A^kx = 0$.

Proof. If $x = 0$ the assertion of the lemma becomes trivial and therefore we assume that $x \neq 0$. Under the conditions of the lemma there is an integer $l \leqslant k$ such that the vectors x, Ax, A^2x, ..., $A^{l-1}x$ are linearly independent while the system of vectors x, Ax, A^2x, ..., $A^{l-1}x$, A^lx is linearly dependent. Then A^lx is expressed linearly in terms of the preceding vectors (why?), that is we have

$$A^lx + a_1A^{l-1}x + \ldots + a_{l-1}Ax + a_lx = 0 \tag{1}$$

where a_1, ..., a_{l-1}, a_l are some scalars. As is known from algebra, every polynomial

$$P(z) = z^l + a_1z^{l-1} + \ldots + a_{l-1}z + a_l$$

can be factored into linear factors (e. g. see [107], Sec. VIII.8):

$$P(z) = (z - z_1)(z - z_2) \ldots (z - z_l) \tag{2}$$

Therefore, by virtue of (1), we can write

$$(A - z_1I)(A - z_2I) \ldots (A - z_lI)x = 0 \tag{3}$$

where I is the identity mapping. The vector $y = (A - z_2I) \ldots$ $\ldots (A - z_lI)x$ is not equal to 0 (why?), and since from (3) we obtain

$$(A - z_1I)y = 0, \text{ i.e. } Ay = z_1y$$

we see that z_1 is an eigenvalue of A and hence, by the hypothesis, $z_1 = 0$. Interchanging the factors in (2) we similarly conclude that z_2, ..., z_l are also eigenvalues of A and $z_2 = \ldots = z_l = 0$. Hence, $P(z) \equiv z^l$ which shows that $a_1 = \ldots = a_{l-1} = a_l = 0$. Thus, it follows from relation (1) that $A^lx = 0$, and therefore we also have $A^kx = A^{k-l}A^lx = 0$, which is what we set out to prove.

For simplicity's sake we shall suppose that $n = 3$, that is the space (R) in which the mapping A is defined is three-dimensional. By the hypothesis, the space (S_1) of the eigenvectors is one-dimensional. Let us now choose in the space (R) a vector $x_0 \overline{\in} (S_1)$ (remember that the symbol $\overline{\in}$ means "does not belong to"); then either $A^2x_0 \neq 0$ or $A^2x_0 = 0$. Let us consider these possible cases.

1. Suppose that $A^2x_0 \neq 0$. Then, according to the above lemma, the vectors $l_1 = x_0$, $l_2 = Ax_0$ and $l_3 = A^2x_0$ are linearly independent and therefore they form a basis in (R). But we have $Al_1 = l_2$, $Al_2 = l_3$ and $Al_3 = 0$ (the latter relation is implied by the same lemma because $Al_3 = A^3x_0$), and consequently the matrix of the

mapping has, in this basis, the form

$$\mathbf{A}' = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Here $\boldsymbol{l}_3$ is an eigenvector, and the subspace spanned by it coincides with (S_1). It can easily be shown (do it!) that the vectors $\boldsymbol{x}$ belonging to the two-dimensional subspace (S_2) spanned by the vectors $\boldsymbol{l}_2$ and $\boldsymbol{l}_3$ are determined by the relation $\mathrm{A}^2\boldsymbol{x} = \boldsymbol{0}$.

2. Now let $\mathrm{A}^2\boldsymbol{x}_0 = \boldsymbol{0}$. Take an arbitrary nonzero eigenvector $\boldsymbol{y}_0 \neq \boldsymbol{0}$ and choose a vector $\boldsymbol{x}_1 \in (R)$ not belonging to the two-dimensional subspace spanned by $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$. Then $\mathrm{A}^2\boldsymbol{x}_1 \neq \boldsymbol{0}$, because, if otherwise, we can write (check it up!) the relations

$$\mathrm{A}(\mathrm{A}\boldsymbol{x}_0) = \boldsymbol{0}, \quad \mathrm{A}\boldsymbol{x}_0 = \alpha\boldsymbol{y}_0 \quad (\alpha \neq 0); \quad \mathrm{A}(\mathrm{A}\boldsymbol{x}_1) = \boldsymbol{0},$$

$$\mathrm{A}\boldsymbol{x}_1 = \beta\boldsymbol{y}_0 \quad (\beta \neq 0),$$

$$\mathrm{A}(\beta\boldsymbol{x}_0 - \alpha\boldsymbol{x}_1) = \beta\alpha\boldsymbol{y}_0 - \alpha\beta\boldsymbol{y}_0 = \boldsymbol{0}, \quad \beta\boldsymbol{x}_0 - \alpha\boldsymbol{x}_1 = \gamma\boldsymbol{y}_0,$$

$$\boldsymbol{x}_1 = \frac{\beta}{\alpha}\boldsymbol{x}_0 - \frac{\gamma}{\alpha}\boldsymbol{y}_0$$

and thus arrive at a contradiction. Hence, if we put $\boldsymbol{l}_1 = \boldsymbol{x}_1$, $\boldsymbol{l}_2 = \mathrm{A}\boldsymbol{x}_1$ and $\boldsymbol{l}_3 = \mathrm{A}^2\boldsymbol{x}_1$ the situation becomes quite the same as in Case 1.

We have considered the case when the eigenvalue λ is equal to zero. To investigate the general case it is sufficient to make use of the following assertion: *an element $\boldsymbol{x}$ is an eigenvector of a mapping* A *corresponding to an eigenvalue* λ *if and only if it is an eigenvector of the mapping* $\mathrm{A} - \alpha\mathrm{I}$ *corresponding to the eigenvalue* $\lambda - \alpha$. Indeed, the relations $\mathrm{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ and $(\mathrm{A} - \alpha\mathrm{I})\boldsymbol{x} = (\lambda - \alpha)\boldsymbol{x}$ are obviously equivalent.

Therefore, if the mapping A in question has an eigenvalue λ the mapping $\mathrm{A}_1 = \mathrm{A} - \lambda\mathrm{I}$ has a zero eigenvalue and thus we arrive at the case that has been already investigated. Let us choose, for the mapping A_1, the vectors $\boldsymbol{l}_1$, $\boldsymbol{l}_2$, $\boldsymbol{l}_3$ as was described above. Then, taking into account that $\mathrm{A}_1 = \mathrm{A} - \lambda\mathrm{I}$, we obtain $(\mathrm{A} - \lambda\mathrm{I})\boldsymbol{l}_1 = \boldsymbol{l}_2$, etc., that is

$$\mathrm{A}\boldsymbol{l}_1 = \lambda\boldsymbol{l}_1 + \boldsymbol{l}_2, \quad \mathrm{A}\boldsymbol{l}_2 = \lambda\boldsymbol{l}_2 + \boldsymbol{l}_3, \quad \mathrm{A}\boldsymbol{l}_3 = \lambda\boldsymbol{l}_3 \tag{4}$$

Consequently, relative to this basis, the matrix of the mapping A has the form

$$\mathbf{A}' = \begin{pmatrix} \lambda & 0 & 0 \\ 1 & \lambda & 0 \\ 0 & 1 & \lambda \end{pmatrix} \tag{5}$$

A matrix of form (5) is called a **Jordan matrix** (a **Jordan submatrix** or a **Jordan block** when it is a part of another matrix) or

a **simple classical matrix**. We have considered the case $n=3$; for $n-1$, $n=2$, $n=4$, etc. Jordan matrices are, respectively, of the form

$$(\lambda),\ \begin{pmatrix}\lambda & 0\\ 1 & \lambda\end{pmatrix},\ \begin{pmatrix}\lambda & 0 & 0 & 0\\ 1 & \lambda & 0 & 0\\ 0 & 1 & \lambda & 0\\ 0 & 0 & 1 & \lambda\end{pmatrix} \text{ etc.}$$

In the general case with an arbitrary n it can be similarly shown that the matrix of a mapping A possessing a single eigenvalue takes the form of a Jordan matrix if we choose the (Jordan) basis $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_n$ where $\boldsymbol{l}_n$ is an eigenvector and the vectors $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots$ are associated with $\boldsymbol{l}_n$ by means of formulas analogous to relations (4) (which specify the vectors $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$ associated with the vector $\boldsymbol{l}_3$ in the case $n=3$). For $n=3$ the subspaces (S_1), (S_2) and $(S_3)=R$ spanned, respectively, by the systems of vectors $\boldsymbol{l}_3$; $\boldsymbol{l}_2, \boldsymbol{l}_3$ and $\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3$ are determined by the relations $(A-\lambda I)\boldsymbol{x}=\boldsymbol{0}$, $(A-\lambda I)^2\boldsymbol{x}=\boldsymbol{0}$ and $(A-\lambda I)^3\boldsymbol{x}=\boldsymbol{0}$, and in the case of an arbitrary n similar relations hold.

2. Mapping with a Single Eigenvalue. Now we proceed to investigate a mapping which, as before, has a single eigenvalue to which there may correspond more than one linearly independent eigenvectors (we remind the reader that every eigenvector is regarded to within an arbitrary nonzero scalar factor). For simplicity, we shall again take the case $n=3$ and consider the eigenvalue $\lambda=0$. The space (S) of the eigenvectors can now be one-, two- or three-dimensional. The case when $\dim(S)=1$ was considered in the foregoing section, and therefore we begin with the two-dimensional space (S). Let us arbitrarily choose a vector $\boldsymbol{l}_1 \overline{\in} (S)$ $(\boldsymbol{l}_1 \in (R))$ and put $\boldsymbol{l}_2 = A\boldsymbol{l}_1$. Then $\boldsymbol{l}_2 \in (S)$ because, if otherwise, we have $A^2\boldsymbol{l}_1 \neq \boldsymbol{0}$, and by the lemma of Sec. 1, the vectors $\boldsymbol{l}_1$, $\boldsymbol{l}_2$ and $A\boldsymbol{l}_2$ can be taken as a basis possessing the same properties as in Sec. 1, which contradicts the condition that (S) is two-dimensional. Let us take an arbitrary vector $\boldsymbol{l}_3 \in (S)$ nonparallel to $\boldsymbol{l}_2$. Then the vectors $\boldsymbol{l}_1$, $\boldsymbol{l}_2$ and $\boldsymbol{l}_3$ form a basis of the space (R) and satisfy the relations

$$A\boldsymbol{l}_1 = \boldsymbol{l}_2, \quad A\boldsymbol{l}_2 = \boldsymbol{0}, \quad A\boldsymbol{l}_3 = \boldsymbol{0}$$

By analogy with Sec. 1, for an arbitrary eigenvalue λ we obtain in this case the relations

$$A\boldsymbol{l}_1 = \lambda\boldsymbol{l}_1 + \boldsymbol{l}_2, \quad A\boldsymbol{l}_2 = \lambda\boldsymbol{l}_2, \quad A\boldsymbol{l}_3 = \lambda\boldsymbol{l}_3$$

We see that $\boldsymbol{l}_2$ and $\boldsymbol{l}_3$ are eigenvectors, and the vector $\boldsymbol{l}_1$ is associated with $\boldsymbol{l}_2$ (in the same manner as in Sec. 1 the vectors $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$ are associated with $\boldsymbol{l}_3$). In the basis $\boldsymbol{l}_1$, $\boldsymbol{l}_2$, $\boldsymbol{l}_3$ thus

constructed the matrix of the mapping takes the quasi-diagonal form

$$\mathbf{A}' = \begin{pmatrix} \lambda & 0 & 0 \\ 1 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}$$

(see Sec. 1.2) having Jordan submatrices of the second and first orders along its principal diagonal. The spaces (S) and (R) are, respectively, determined by the relations $(\mathrm{A}-\lambda\mathrm{I})\boldsymbol{x}=\boldsymbol{0}$ and $(\mathrm{A}-\lambda\mathrm{I})^2\boldsymbol{x}=\boldsymbol{0}$.

Finally, let us turn to the case when the space (S) is three-dimensional, i.e. $(S)=(R)$. Then we can take an arbitrary basis $\boldsymbol{l}_1$, $\boldsymbol{l}_2$, $\boldsymbol{l}_3$ and the matrix of the mapping has the form

$$\mathbf{A}' = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}$$

Thus we obtain a diagonal matrix which can be regarded as a quasi-diagonal matrix with three first-order Jordan submatrices on the principal diagonal. In this case all the vectors of (R) satisfy the relation $(\mathrm{A}-\lambda\mathrm{I})\boldsymbol{x}=\boldsymbol{0}$.

It turns out that this result can be extended to a space (R) of an arbitrary dimension p if the mapping A possesses a single eigenvalue λ. Namely, in an appropriately chosen basis $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_p$ the matrix $\mathbf{A}'$ of the mapping takes a quasi-diagonal form with Jordan blocks of orders $p_1, p_2, \ldots, p_d$ along its principal diagonal where p_j $(j=1, 2, \ldots, d)$ are some positive integers satisfying the relation $p_1+p_2+\ldots+p_d=p$. Each of these Jordan submatrices has all elements of its principal diagonal equal to λ. The vectors $\boldsymbol{l}_{p_1}$, $\boldsymbol{l}_{p_1+p_2}$ etc. in this basis are eigenvectors, the vectors $\boldsymbol{l}_1$, $\boldsymbol{l}_2, \ldots, \boldsymbol{l}_{p_1-1}$ are associated with $\boldsymbol{l}_{p_1}$ (in the same way as in Sec. 1 the vectors $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$ are associated with $\boldsymbol{l}_3$), the vectors $\boldsymbol{l}_{p_1+1}$, $\boldsymbol{l}_{p_1+2}, \ldots, \boldsymbol{l}_{p_1+p_2-1}$ are associated with $\boldsymbol{l}_{p_1+p_2}$ and so on. The d-dimensional subspace (S_1) spanned by the vectors $\boldsymbol{l}_{p_1}, \boldsymbol{l}_{p_1+p_2}, \ldots, \boldsymbol{l}_{p_1+p_2+\ldots+p_d=p}$ is nothing but the subspace of the eigenvectors, that is it consists of all the solutions of the equation $(\mathrm{A}-\lambda\mathrm{I})\boldsymbol{x}=\boldsymbol{0}$. The subspace (S_2) spanned by all eigenvectors and the vectors $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_{p_1+p_2-1}$ coincides with the subspace of solutions of the equation $(\mathrm{A}-\lambda\mathrm{I})^2\boldsymbol{x}=\boldsymbol{0}$, etc.

Let d_k be the dimension of the subspace (S_k) (we put $d_1=d$). The dimension of (S_2) is greater by d_2-d_1 than that of (S_1). But this difference of the dimensions must be equal to the number of associated vectors added to the eigenvectors, that is to the number of the Jordan submatrices whose order is greater than unity (why?).

Thus, the number of the submatrices of order unity is equal to $d_1-(d_2-d_1)=2d_1-d_2$. We similarly conclude that the number of the Jordan submatrices of order two is equal to $(d_2-d_1)-(d_3-d_2)=2d_2-d_1-d_3$, etc. (Let the reader check that the sum of the expressions thus obtained is equal to the total number $d_1=d$ of all the submatrices.)

3. General Case. Let there be given an arbitrary linear mapping A of an n-dimensional complex space (R) into itself. A vector $\boldsymbol{x}\in(R)$ satisfying the relation $(\mathrm{A}-\lambda\mathrm{I})^k\boldsymbol{x}=\boldsymbol{0}$ for certain k ($k=1, 2, 3, \ldots$) is said to be a **principal vector of grade k belonging to the given value of λ**. The properties below are easily proved.

1. *The set (R') of all principal vectors belonging to a given value λ forms an invariant subspace of (R).* Indeed, if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are two arbitrary vectors of this kind, that is if $(\mathrm{A}-\lambda\mathrm{I})^{k_1}\boldsymbol{x}_1=\boldsymbol{0}$ and $(\mathrm{A}-\lambda\mathrm{I})^{k_2}\boldsymbol{x}_2=\boldsymbol{0}$ (for definiteness, let $k_1\geqslant k_2$), and $\boldsymbol{x}=\alpha\boldsymbol{x}_1+\beta\boldsymbol{x}_2$ then we have the relation $(\mathrm{A}-\lambda\mathrm{I})^{k_1}\boldsymbol{x}=\boldsymbol{0}$ (why?), which means that $\boldsymbol{x}$ also belongs to (R'). Consequently, (R') is a subspace of (R). The invariance of (R') under the mapping A directly follows from the identity $(\mathrm{A}-\lambda\mathrm{I})^k(\mathrm{A}\boldsymbol{x})\equiv\mathrm{A}[(\mathrm{A}-\lambda\mathrm{I})^k\boldsymbol{x}]$ (let the reader carefully examine this argument!).

2. *The dimension of (R') is nonzero if and only if λ is an eigenvalue of the mapping* A. Indeed, if λ is an eigenvalue, the space (R') must contain the corresponding subspace of the eigenvectors (for which $(\mathrm{A}-\lambda\mathrm{I})\boldsymbol{x}=\boldsymbol{0}$). Conversely, let $\dim(R')\neq 0$, i.e. let there exist at least one principal vector $\boldsymbol{x}\neq\boldsymbol{0}$ for the given λ. Taking the least exponent k for which $(\mathrm{A}-\lambda\mathrm{I})^k\boldsymbol{x}=\boldsymbol{0}$ and putting $(\mathrm{A}-\lambda\mathrm{I})^{k-1}\boldsymbol{x}=\boldsymbol{y}$ we see that $\boldsymbol{y}\neq\boldsymbol{0}$ and $(\mathrm{A}-\lambda\mathrm{I})\boldsymbol{y}=\boldsymbol{0}$, and hence λ is an eigenvalue of the mapping A.

Taking into account this property, we shall only consider principal vectors belonging to eigenvalues of the mapping A. Let $\lambda_1, \lambda_2, \ldots, \lambda_m$ designate all the *pairwise distinct* eigenvalues and $(R_1), (R_2), \ldots, (R_m)$ the corresponding subspaces of principal vectors.

3. *The space (R) is representable as the direct sum of the subspaces $(R_1), (R_2), \ldots, (R_m)$.*

We shall split the proof of this property into two stages. To begin with, we shall prove the **Hamilton-Cayley theorem** which states that *if $P(\lambda)=\det(\mathbf{A}-\lambda\mathbf{I})$ is the characteristic polynomial of a matrix* A *then $P(\mathbf{A})=0$*, or, in other words, *every square matrix satisfies its characteristic equation*. Let us denote by $\mathbf{Q}(\lambda)$ the transpose of the matrix composed of the algebraic adjuncts (cofactors) of the elements of the determinant $\det(\mathbf{A}-\lambda\mathbf{I})$. The elements of the matrix $\mathbf{Q}(\lambda)$ (and of $\mathbf{A}-\lambda\mathbf{I}$ as well) are polynomials in λ. Such matrices are termed **polynomial matrices** or **λ-matrices**. For the matrices $\mathbf{Q}(\lambda)$ and $\mathbf{A}-\lambda\mathbf{I}$ we have the relation

$$(\mathbf{A}-\lambda\mathbf{I})\,\mathbf{Q}(\lambda)=P(\lambda)\,\mathbf{I} \tag{6}$$

(why?). On the other hand, it is readily proved that any polynomial of the form $P(u)-P(v)$ is divisible by $u-v$, that is $P(u)-P(v)\equiv(u-v)\Phi(u, v)$ where Φ is a polynomial in two variables u and v. It follows from (6) and the last assertion that

$$P(\mathbf{A})=P(\lambda\mathbf{I})+(\mathbf{A}-\lambda\mathbf{I})\Phi(\mathbf{A}, \lambda\mathbf{I})=P(\lambda)\mathbf{I}+(\mathbf{A}-\lambda\mathbf{I})\Phi(\mathbf{A}, \lambda\mathbf{I})=$$
$$=(\mathbf{A}-\lambda\mathbf{I})[\mathbf{Q}(\lambda)+\Phi(\mathbf{A}, \lambda\mathbf{I})]=\mathbf{A}[\mathbf{Q}(\lambda)+\Phi(\mathbf{A}, \lambda\mathbf{I})]-$$
$$-\lambda[\mathbf{Q}(\lambda)+\Phi(\mathbf{A}, \lambda\mathbf{I})]$$

(let the reader carefully perform all these operations on matrices). The expression in the rightmost square brackets must be identically (with respect to λ) equal to zero because, if otherwise, its element containing the highest power of λ multiplied by λ would not be cancelled out in the right-hand side and therefore the right member of the last relation would be dependent on λ while the left member $P(\mathbf{A})$ would not involve λ. This argument completes the proof of the Hamilton-Cayley theorem.

Now, proceeding to prove Property 3, we shall suppose, for simplicity's sake, that the mapping A has only two distinct eigenvalues λ_1 and λ_2. Let $P(\lambda)$ be the characteristic polynomial of this mapping, that is the characteristic polynomial of the matrix $\mathbf{A}$ of the mapping A, which, as is known, does not depend on the particular choice of the basis. Under the hypothesis assumed, we can write $P(\lambda)\equiv(\lambda_1-\lambda)^{n_1}(\lambda_2-\lambda)^{n_2}$. Let us temporarily denote by (R_j) the subspace of vectors for which $(\mathrm{A}-\lambda_j\mathrm{I})^{n_j}\boldsymbol{x}=\boldsymbol{0}$.

Taking the decomposition of the rational fraction $\frac{1}{P(\lambda)}$ into partial rational fractions of the first type (e.g. see [107], Sec. VIII.10) and multiplying both sides by $P(\lambda)$ we arrive at the identity

$$1\equiv D_1(\lambda)(\lambda-\lambda_1)^{n_1}+D_2(\lambda)(\lambda-\lambda_2)^{n_2}$$

where $D_j(\lambda)$ are some polynomials. It follows that

$$\mathrm{I}\equiv D_1(\mathrm{A})(\mathrm{A}-\lambda_1\mathrm{I})^{n_1}+D_2(\mathrm{A})(\mathrm{A}-\lambda_2\mathrm{I})^{n_2} \tag{7}$$

By (7), we can write, for any vector $\boldsymbol{x}$, the relation

$$\boldsymbol{x}=D_2(\mathrm{A})(\mathrm{A}-\lambda_2\mathrm{I})^{n_2}\boldsymbol{x}+D_1(\mathrm{A})(\mathrm{A}-\lambda_1\mathrm{I})^{n_1}\boldsymbol{x} \tag{8}$$

But the Hamilton-Cayley theorem implies that the first term of the right-hand side belongs to (R_1) and the second to (R_2) (why?). Consequently, to prove the validity of the desired direct decomposition it only remains to show that the representation of every vector $\boldsymbol{x}$ as a sum of vectors belonging to (R_1) and (R_2) is unique. But, for this purpose, it is obviously sufficient to prove that the zero vector can be uniquely expressed in that form. Therefore, let us suppose that

$$\boldsymbol{0}=\boldsymbol{x}_1+\boldsymbol{x}_2 \text{ where } \boldsymbol{x}_1\in(R_1),\ \boldsymbol{x}_2\in(R_2)$$

Applying the mapping $D_1(A)(A-\lambda_1 I)^{n_1}$ to both sides of this relation and taking advantage of identity (7) and of the definition of the spaces (R_j) we obtain

$$\boldsymbol{0} = D_1(A)(A-\lambda_1 I)^{n_1}\boldsymbol{x}_2 = \boldsymbol{x}_2 - D_2(A)(A-\lambda_2 I)^{n_2}\boldsymbol{x}_2 = \boldsymbol{x}_2$$

and, similarly, $\boldsymbol{x}_1 = \boldsymbol{0}$, which completes the proof. Thus, the validity of the decomposition of (R) into the direct sum of (R_1) and (R_2) has been proved.

It now remains to check that the temporary definition of (R_j) which we have used is equivalent to the original definition, i.e. if $(A-\lambda_1 I)^k \boldsymbol{x} = \boldsymbol{0}$ for an integer k then we also have $(A-\lambda_1 I)^{n_1}\boldsymbol{x} = \boldsymbol{0}$. In fact, let us apply the mapping on the left-hand side of (7) to the left member of relation (8) and the mapping on the right-hand side of (7) to the right member of (8) a sufficient number of times. Then the Hamilton-Cayley theorem indicates that we shall arrive at the equality $\boldsymbol{x} = [D_2(A)]^r (A-\lambda_2 I)^{n_2 r}\boldsymbol{x}$ which implies the desired assertion. Thus, Property 3 has been proved.

We see that the problem of describing the structure of the mapping A in the space (R) reduces to the analogous question for each of the subspaces (R_j). The latter can be answered if we take advantage of the following property:

4. *The mapping* A *regarded as defined in the subspace* (R_j) *possesses only one eigenvalue* λ_j. Virtually, let $\boldsymbol{x} \in (R_j)$, $\boldsymbol{x} \neq \boldsymbol{0}$ and $A\boldsymbol{x} = \lambda \boldsymbol{x}$. According to the definition of (R_j), we have $(A-\lambda_j I)^k \boldsymbol{x} = \boldsymbol{0}$. Opening the brackets on the left-hand side and taking into account that $A\boldsymbol{x} = \lambda \boldsymbol{x}$, $A^2\boldsymbol{x} = A(A\boldsymbol{x}) = A(\lambda \boldsymbol{x}) = \lambda^2 \boldsymbol{x}$, $A^3\boldsymbol{x} = \lambda^3 \boldsymbol{x}$, etc., we derive (check it up!) the equality $(\lambda-\lambda_j)^k \boldsymbol{x} = \boldsymbol{0}$ whence $\lambda = \lambda_j$.

Now we see that the results of Secs. 1.2 and 2 indicate that, after the corresponding basis consisting of the eigenvectors and associated vectors has been chosen in each subspace (R_j), the matrix of the mapping A takes a quasi-diagonal form with Jordan blocks along the principal diagonal. A quasi-diagonal matrix of this type is said to be in **classical canonical form** or in **Jordan normal** (or **canonical**) **form**. Every Jordan submatrix entering into such a matrix (called a **classical canonical matrix**) contains one of the eigenvalues of the corresponding mapping A on its principal diagonal; the number of the Jordan submatrices corresponding to each eigenvalue is equal to the number of linearly independent eigenvectors associated with this eigenvalue. The orders of the Jordan submatrices can be determined as described at the end of Sec. 2, that is can be expressed in terms of the dimensions of the subspaces of vectors which satisfy, for the given j, the relations $(A-\lambda_j I)\boldsymbol{x} = \boldsymbol{0}$, $(A-\lambda_j I)^2 \boldsymbol{x} = \boldsymbol{0}$, etc. The order in which the constituent Jordan submatrices are arranged along the principal diagonal is inessential because we can always renumber the invariant subspaces, which results in the corresponding permutation of

the submatrices. The exact type of the given classical canonical matrix is thus specified by the so-called **Segre characteristic** which is the set of integers indicating the orders of the Jordan submatrices, those integers which correspond to submatrices containing the same characteristic root (eigenvalue) being bracketed together.

The above result can be formulated purely in terms of matrix algebra. Consider an arbitrary complex square matrix **A** of order n. It can be interpreted as the matrix of a mapping A of an n-dimensional complex linear space into itself, with respect to an arbitrarily taken basis. If we pass to another basis the matrix of the mapping is transformed according to the formula $\mathbf{A}' = \mathbf{H}^{-1}\mathbf{A}\mathbf{H}$ (e.g. see [107], Sec. XI.7) where **H** is an invertible (nondegenerate) transformation matrix from the former basis to the latter. Thus we conclude that *for every matrix* **A** *it is possible to construct a nonsingular matrix* (i.e. a *nondegenerate matrix possessing the inverse matrix*) **H** *such that the matrix* $\mathbf{A}' = \mathbf{H}^{-1}\mathbf{A}\mathbf{H}$ *has the Jordan normal form**). The elements of the principal diagonal of every Jordan submatrix of the matrix $\mathbf{A}'$ are equal to one of the eigenvalues of the matrix **A**. The orders of the submatrices are determined as described above. It should also be taken into account that the dimension of the subspace of vectors satisfying an equation of the form $(\mathrm{A}-\lambda_j \mathrm{I})^k \boldsymbol{x} = \boldsymbol{0}$ is equal to $n-\operatorname{rank}(\mathbf{A}-\lambda_j \mathbf{I})^k$ (e.g. see [107], Secs. XI.5 and 6), and consequently the orders of the constituent submatrices are also expressible in terms of the properties of the original matrix **A**.

There is another method of determining the orders of the Jordan submatrices. We shall not go into particulars here (for more detail we refer the reader to [39], [47] and [118]) and only mention that this method associates, with a given matrix **A**, a sequence of polynomials of the form $(\lambda-\lambda_j)^k$. These polynomials (called **elementary divisors** of the matrix **A**) are found according to a rather complicated rule and possess the property that, after the matrix **A** has been reduced to the corresponding classical canonical matrix $\mathbf{A}'$, to each of these divisors there corresponds, in $\mathbf{A}'$, a Jordan submatrix of order k with the elements of its principal diagonal equal to the number λ_j.

4. Mapping of a Real Space. Now let A be a linear mapping of a real space (R) into itself. Then if the characteristic equation determining the eigenvalues has only real roots all the foregoing considerations remain valid, that is the matrix of the mapping A has a classical canonical form in an appropriately chosen basis. It is obvious that to every Jordan submatrix there corresponds an invariant subspace whose dimension is equal to the order of

*) Two matrices **A** and **B** satisfying the relation $\mathbf{A} = \mathbf{H}^{-1}\mathbf{B}\mathbf{H}$ (or, equivalently, $\mathbf{B} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$, $\mathbf{P} = \mathbf{H}^{-1}$) where **H** is a nonsingular matrix are said to be **similar**. Thus, *every matrix is similar to a Jordan canonical matrix.—Tr.*

the submatrix, and the whole space (R) is the direct sum of these subspaces. In the general case this direct decomposition is composed of "smaller" subspaces than those used in Sec. 3 and denoted by (R_j). Let us agree that now the symbols (R_j) designate these smaller invariant subspaces corresponding to the Jordan submatrices. Therefore now, to different j, there may correspond equal values of λ_j, and the space (R) is represented as the direct sum of the subspaces (R_j) in each of which the mapping A has a matrix of the form of a simple classical matrix (Jordan submatrix) with the value λ_j on its principal diagonal, after an appropriate basis has been chosen.

If a subspace of the type (R_j) is one-dimensional the action of the mapping A in this subspace reduces to λ_j-fold stretching. In particular, this is the case when λ_j is a simple eigenvalue and also in a more general case when the eigenvalue λ_j is multiple but the maximum number of linearly independent eigenvectors corresponding to λ_j is equal to its multiplicity. (According to Sec. 3, this number is always equal to or less than the multiplicity.) If (R_j) is a two-dimensional subspace then it is possible to choose a basis (to which some coordinate axes x_1 and x_2 correspond) in such a way that, in (R_j), the mapping can be written, in coordinate form, as

$$y_1 = \lambda_j x_1, \quad y_2 = x_1 + \lambda_j x_2$$

where (x_1, x_2) are the coordinates of a given vector (a preimage) and (y_1, y_2) are the coordinates of the image under the mapping. If $\lambda_j \neq 0$ these formulas determine a combination of a uniform (λ_j-fold) magnification and a shift along the x_2-axis. Let the reader describe this mapping for the case $\lambda_j = 0$ and also consider the geometric interpretation of a mapping corresponding to a Jordan submatrix of the third order.

Now let us proceed to consider the case when the characteristic equation may have imaginary (nonreal) roots. A root $\lambda = \mu + i\nu$ (where μ and ν are real and $\nu \neq 0$) is not an eigenvalue of the mapping A in the real space (R). But we can take a basis $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_n$ in the space (R) and extend (R) to a complex linear space (S) in such a way that this system of vectors also serves as a basis of (S). Indeed, for this purpose it is sufficient to introduce as elements of (S) all the possible formal linear combinations $\alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 + \ldots + \alpha_n\boldsymbol{x}_n$ (where α_k's are arbitrary complex numbers) and define addition and multiplication by complex scalars in a natural manner by putting

$$(\alpha_1\boldsymbol{x}_1 + \ldots + \alpha_n\boldsymbol{x}_n) + (\alpha_1'\boldsymbol{x}_1 + \ldots + \alpha_n'\boldsymbol{x}_n) = \\ = (\alpha_1 + \alpha_1')\boldsymbol{x}_1 + \ldots + (\alpha_n + \alpha_n')\boldsymbol{x}_n$$

and

$$\beta(\alpha_1\boldsymbol{x}_1 + \ldots + \alpha_n\boldsymbol{x}_n) = (\alpha_1\beta)\boldsymbol{x}_1 + \ldots + (\alpha_n\beta)\boldsymbol{x}_n$$

The mapping A can also be extended to a linear mapping of (S) into itself because, as is known, every linear mapping is completely characterized if we know the images of the base vectors. The matrix of the extended mapping coincides in the basis $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ with the matrix of the original mapping, and therefore the root $\lambda = \mu + i\nu$ is an eigenvalue of the extended mapping as well.

It is known from algebra that every algebraic equation with real coefficients having an imaginary root always possesses the conjugate complex root of the same multiplicity (e.g. see [107], Sec. VIII.8). Therefore, the number $\lambda^* = \mu - i\nu$ is also an eigenvalue of the extended mapping A. We thus can write down relations of type (4), in the coordinates relative to the basis $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, and then pass to the complex conjugates and thus conclude that if, for instance, $\boldsymbol{l}_3$ is an eigenvector corresponding to the eigenvalue λ and $\boldsymbol{l}_1$, $\boldsymbol{l}_2$ are the associated vectors then the vectors $\boldsymbol{l}_3^*$, $\boldsymbol{l}_1^*$ and $\boldsymbol{l}_2^*$ with the conjugate complex coordinates are, respectively, an eigenvector for the eigenvalue λ^* and the vectors associated with it (the same conclusion can also be drawn in the general case). These vectors do not belong to (R) because their coordinates are imaginary. Separating the real and imaginary parts of the coordinates of these vectors we can form the corresponding vectors $\boldsymbol{l}_k'$ and $\boldsymbol{l}_k''$ $(k = 1, 2, 3)$ belonging to (R) for which

$$\boldsymbol{l}_k = \boldsymbol{l}_k' + i\boldsymbol{l}_k'', \quad \boldsymbol{l}_k^* = \boldsymbol{l}_k' - i\boldsymbol{l}_k'' \qquad (k = 1, 2, 3)$$

Substituting the new vectors into (4) we obtain (check it up!) the relations

$$\mathrm{A}\boldsymbol{l}_k' = \mu\boldsymbol{l}_k' - \nu\boldsymbol{l}_k'' + \boldsymbol{l}_{k+1}', \quad \mathrm{A}\boldsymbol{l}_k'' = \nu\boldsymbol{l}_k' + \mu\boldsymbol{l}_k'' + \boldsymbol{l}_{k+1}'' \qquad (k = 1, 2)$$

$$\mathrm{A}\boldsymbol{l}_3' = \mu\boldsymbol{l}_3' - \nu\boldsymbol{l}_3'', \quad \mathrm{A}\boldsymbol{l}_3'' = \nu\boldsymbol{l}_3' + \mu\boldsymbol{l}_3''$$

(The real parts of the coordinates of the vectors $\boldsymbol{l}_k^*$ coincide with those of the vectors $\boldsymbol{l}_k$ and their imaginary parts only differ in sign from those of $\boldsymbol{l}_k$, and consequently the above operation performed on $\boldsymbol{l}_k^*$ does not yield any new relations.) Therefore, if we replace, in the basis relative to which the matrix of the mapping A (considered in (S)) takes the Jordan normal form, the six vectors $\boldsymbol{l}_k$, $\boldsymbol{l}_k^*$ $(k = 1, 2, 3)$ by the six vectors $\boldsymbol{l}_k'$, $\boldsymbol{l}_k''$ $(k = 1, 2, 3)$ and consider the invariant subspace of (R) spanned by these vectors, the mapping A will have in the new basis the matrix

$$\begin{pmatrix} \mu & \nu & 0 & 0 & 0 & 0 \\ -\nu & \mu & 0 & 0 & 0 & 0 \\ 1 & 0 & \mu & \nu & 0 & 0 \\ 0 & 1 & -\nu & \mu & 0 & 0 \\ 0 & 0 & 1 & 0 & \mu & \nu \\ 0 & 0 & 0 & 1 & -\nu & \mu \end{pmatrix} \tag{9}$$

A transformation of that kind can be applied to all imaginary eigenvalues, and this results in the construction of a basis in the original real space (R) relative to which the matrix of the mapping A (again considered as acting in (R)) has a quasi-diagonal form with real submatrices of type (5) (of any order) and of type (9) (of any even order) along its principal diagonal. This result can be similarly expressed in terms of real matrix algebra, i.e. can be formulated as a theorem on reduction of a real matrix to its **real canonical form** (of the type described above), the transformation matrix **H** (entering into the formulation of the theorem) also being real.

In the simplest case matrix (9) has the form

$$\begin{pmatrix} \mu & \nu \\ -\nu & \mu \end{pmatrix}$$

Putting $\mu = M\cos\alpha$ and $-\nu = M\sin\alpha$ $(M > 0)$ we can regard l_1', l_1'' as a Cartesian basis in a plane and interpret (this can easily be justified) the action of the mapping in this plane as the combination of the uniform M-fold magnification and the rotation through the angle α.

5. Application to Constructing Functions of Matrices. Let $f(x)$ be a scalar function possessing the series expansion

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + \ldots$$

and **A** be a square matrix. Then we define the value of the function $f(\mathbf{A})$ as a matrix determined by the formula

$$f(\mathbf{A}) = a_0\mathbf{I} + a_1\mathbf{A} + a_2\mathbf{A}^2 + \ldots + a_n\mathbf{A}^n + \ldots \qquad (10)$$

According to Sec. 3, we can write $\mathbf{A} = \mathbf{H}\mathbf{A}'\mathbf{H}^{-1}$ where $\mathbf{A}'$ is the corresponding classical canonical matrix; then $\mathbf{A}^2 = \mathbf{H}\mathbf{A}'\mathbf{H}^{-1}\mathbf{H}\mathbf{A}'\mathbf{H}^{-1} = \mathbf{H}\mathbf{A}'^2\mathbf{H}^{-1}$, $\mathbf{A}^3 = \mathbf{H}\mathbf{A}'^3\mathbf{H}^{-1}$, etc., and thus we obtain

$$f(\mathbf{A}) = \mathbf{H}a_0\mathbf{I}\mathbf{H}^{-1} + \mathbf{H}a_1\mathbf{A}'\mathbf{H}^{-1} + \mathbf{H}a_2\mathbf{A}'^2\mathbf{H}^{-1} + \ldots = \mathbf{H}f(\mathbf{A}')\mathbf{H}^{-1}$$

We shall compute $f(\mathbf{A}')$ on the basis of the following property which can readily be proved: *the multiplication of two quasi-diagonal matrices of the same type (i.e. with square submatrices of the same size arranged in the same order along the principal diagonals) results in a quasi-diagonal matrix of that type which is found according to the formula*

$$\begin{pmatrix} \mathbf{A}_1 & & & 0 \\ & \mathbf{A}_2 & & \\ & & \ddots & \\ 0 & & & \mathbf{A}_m \end{pmatrix} \begin{pmatrix} \mathbf{B}_1 & & & 0 \\ & \mathbf{B}_2 & & \\ & & \ddots & \\ 0 & & & \mathbf{B}_m \end{pmatrix} = \begin{pmatrix} \mathbf{A}_1\mathbf{B}_1 & & & 0 \\ & \mathbf{A}_2\mathbf{B}_2 & & \\ & & \ddots & \\ 0 & & & \mathbf{A}_m\mathbf{B}_m \end{pmatrix}$$

which means that the submatrices of the same size occupying the same position on the diagonal are multiplied independently. It follows, in particular, that if

$$\mathbf{A}'=\begin{pmatrix}\mathbf{A}_1 & & & \\ & \mathbf{A}_2 & & 0 \\ & & \ddots & \\ & 0 & & \mathbf{A}_m\end{pmatrix} \text{ then } f(\mathbf{A}')=\begin{pmatrix}f(\mathbf{A}_1) & & & \\ & f(\mathbf{A}_2) & & 0 \\ & & \ddots & \\ & 0 & & f(\mathbf{A}_m)\end{pmatrix}$$

(why?). Therefore, it now remains to compute $f(\mathbf{A}_k)$ where $\mathbf{A}_k$ is a simple classical matrix (a Jordan submatrix). For definiteness, let us consider a Jordan submatrix of form (5). It is readily found (prove it!) that its subsequent powers $\mathbf{A}_k^2$, $\mathbf{A}_k^3$, $\mathbf{A}_k^4$, etc., are, respectively, equal to

$$\begin{pmatrix}\lambda^2 & 0 & 0\\ 2\lambda & \lambda^2 & 0\\ 1 & 2\lambda & \lambda^2\end{pmatrix},\quad \begin{pmatrix}\lambda^3 & 0 & 0\\ 3\lambda^2 & \lambda^3 & 0\\ 3\lambda & 3\lambda^2 & \lambda^3\end{pmatrix},\quad \begin{pmatrix}\lambda^4 & 0 & 0\\ 4\lambda^3 & \lambda^4 & 0\\ 6\lambda^2 & 4\lambda^3 & \lambda^4\end{pmatrix} \text{ etc.}$$

We see that the coefficients in powers of λ entering into these matrices are equal to those obtained when we write the expressions $(\lambda+1)^2$, $(\lambda+1)^3$, $(\lambda+1)^4$, etc., as polynomials in descending (or ascending) powers of λ. Consequently, the general formula is written as

$$\begin{pmatrix}\lambda & 0 & 0\\ 1 & \lambda & 0\\ 0 & 1 & \lambda\end{pmatrix}^n=\begin{pmatrix}\lambda^n & 0 & 0\\ \binom{n}{1}\lambda^{n-1} & \lambda^n & 0\\ \binom{n}{2}\lambda^{n-2} & \binom{n}{1}\lambda^{n-1} & \lambda^n\end{pmatrix}$$

where $\binom{n}{k}=\frac{n(n-1)\ldots(n-k+1)}{k!}$ is a binomial coefficient. Substituting this result into (10) and taking into account the relations

$$a_0+a_1\lambda+a_2\lambda^2+a_3\lambda^3+\ldots=f(\lambda)$$

$$a_1+2a_2\lambda+3a_3\lambda^2+4a_4\lambda^3+\ldots=f'(\lambda)=\frac{1}{1!}f'(\lambda)$$

$$\binom{2}{2}a_2+\binom{3}{2}a_3\lambda+\binom{4}{2}a_4\lambda^2+\ldots=$$

$$=\frac{1}{2!}(2\cdot 1a_2+3\cdot 2a_3\lambda+4\cdot 3a_4\lambda^2+\ldots)=\frac{1}{2!}f''(\lambda)$$

we finally conclude that

$$\text{if } \mathbf{A}_k=\begin{pmatrix}\lambda & 0 & 0\\ 1 & \lambda & 0\\ 0 & 1 & \lambda\end{pmatrix} \text{ then } f(\mathbf{A}_k)=\begin{pmatrix}f(\lambda) & 0 & 0\\ \frac{1}{1!}f'(\lambda) & f(\lambda) & 0\\ \frac{1}{2!}f''(\lambda) & \frac{1}{1!}f'(\lambda) & f(\lambda)\end{pmatrix} \quad (11)$$

The same result is obtained for Jordan matrices of any order, which makes it possible to complete the computation of $f(\mathbf{A})$.

As an example, let us consider the notion of the **logarithm of a matrix A**, i.e. investigate the problem of constructing a matrix **B** such that $e^{\mathbf{B}} = \mathbf{A}$. It turns out that the logarithm always exists when the eigenvalues of the matrix **A** are all nonzero. As above, to establish this result it is sufficient to find the logarithm of a simple classical matrix of form (5) for $\lambda \neq 0$. Let us put

$$\mathbf{A}_k' = \begin{pmatrix} \lambda & 0 & 0 \\ 1 & \lambda & 0 \\ 0 & 1 & \lambda \end{pmatrix}, \quad \mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ \lambda^{-1} & 1 & 0 \\ 0 & \lambda^{-1} & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{\Lambda} = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}$$

Then it is readily seen (let the reader carry out the calculations!) that $\mathbf{A}_k' = \mathbf{\Lambda P}$, and therefore

$$\ln \mathbf{A}_k' = \ln (\mathbf{\Lambda P}) = (\operatorname{Ln} \lambda)\, \mathbf{I} + \ln (\mathbf{I} + \mathbf{P}) \tag{12}$$

where $\operatorname{Ln} \lambda$ is the general expression of the infinite-valued logarithm. It should be noted that the ordinary rule of finding the logarithm of a product of numbers as the sum of the logarithms of the factors is by far not always applicable to matrices because, for matrices, this rule is based on the equality $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}} e^{\mathbf{B}}$ (equivalent to $\ln (\mathbf{KL}) = \ln \mathbf{K} + \ln \mathbf{L}$ for $\mathbf{K} = e^{\mathbf{A}}$, $\mathbf{L} = e^{\mathbf{B}}$) which is only valid for *commuting matrices* **A** and **B** for which $\mathbf{AB} = \mathbf{BA}$ (e.g. see [107], Sec. XVII.18). The second summand on the right-hand side of (12) can be computed by means of the expansion $\ln (1 + x) = \frac{x}{1} - \frac{x^2}{2} + \ldots$ whose right-hand side involves only a finite number of nonzero terms after **I** and **P** have been, respectively, substituted for 1 and x (why?).

On more detail concerning the properties of matrices and matrix functions and also the solution of more complicated matrix equations we refer the reader to [44].

In applications we sometimes deal with functions of the form $f(t\mathbf{A})$ where t is a parameter. Then if, for instance, $\mathbf{A} = \begin{pmatrix} \lambda & 0 & 0 \\ 1 & \lambda & 0 \\ 0 & 1 & \lambda \end{pmatrix}$ we resort, instead of (11), to the formula

$$f(t\mathbf{A}) = \begin{pmatrix} f(t\lambda) & 0 & 0 \\ \frac{t}{1!} f'(t\lambda) & f(t\lambda) & 0 \\ \frac{t^2}{2!} f''(t\lambda) & \frac{t}{1!} f'(t\lambda) & f(t\lambda) \end{pmatrix} \tag{13}$$

whose deduction is left to the reader. For instance, the application of this formula shows that

$$\exp\begin{pmatrix} t\lambda & 0 & 0 \\ t & t\lambda & 0 \\ 0 & t & t\lambda \end{pmatrix} = \begin{pmatrix} e^{t\lambda} & 0 & 0 \\ \frac{t}{1!}e^{t\lambda} & e^{t\lambda} & 0 \\ \frac{t^2}{2!}e^{t\lambda} & \frac{t}{1!}e^{t\lambda} & e^{t\lambda} \end{pmatrix} = e^{t\lambda}\begin{pmatrix} 1 & 0 & 0 \\ \frac{t}{1!} & 1 & 0 \\ \frac{t^2}{2!} & \frac{t}{1!} & 1 \end{pmatrix}$$

6. Another Representation of a Mapping of a Real Space. Consider an arbitrary linear mapping A of a real Euclidean space (R) into itself. It was shown in Sec. 1.6 that if the mapping A is self-adjoint it simply reduces to a combination of uniform stretchings along the mutually perpendicular directions of the eigenvectors, the magnification coefficients being equal to the corresponding eigenvalues. In this section we shall prove that *a mapping* A *of the general type is a combination of a self-adjoint mapping and an orthogonal mapping* (which reduces to a rotation of (R) about the origin) *and that the squares of the magnification coefficients are equal to the corresponding eigenvalues of the self-adjoint mapping* AA^*.

This proposition can also be formulated solely in terms of matrix algebra. Take an arbitrary real square matrix $\mathbf{A}$ and consider it as the matrix of a mapping A in a Euclidean basis of a Euclidean space. Then, if the above assertion is true, we obtain, after passing to the basis consisting of the eigenvectors of the above-mentioned self-adjoint mapping AA^*, the relation

$$\mathbf{H}^{-1}\mathbf{A}\mathbf{H} = \mathbf{U}\Lambda \text{ whence } \mathbf{A} = \mathbf{H}\mathbf{U}\Lambda\mathbf{H}^{-1} = \mathbf{K}\Lambda\mathbf{L} \qquad (14)$$

where $\mathbf{K} = \mathbf{H}\mathbf{U}$, $\mathbf{L} = \mathbf{H}^{-1}$, Λ is a real diagonal matrix, and $\mathbf{K}$ and $\mathbf{L}$ are orthogonal matrices because such are the matrices $\mathbf{H}$ and $\mathbf{U}$. The matrix Λ^2 is then diagonal, and the elements of its principal diagonal are equal to the eigenvalues of the matrix $\mathbf{A}\mathbf{A}^*$. (Let the reader show that the latter matrix is symmetric and all its eigenvalues are nonnegative.) Conversely, if representation (14) is valid we can rewrite it in the form $\mathbf{A} = (\mathbf{K}\mathbf{L})(\mathbf{L}^{-1}\Lambda\mathbf{L}) = (\mathbf{K}\mathbf{L})(\mathbf{L}^*\Lambda\mathbf{L})$ which is equivalent (why?) to the above assertion concerning the structure of the mapping A.

Thus, we must prove formula (14). To this end, let us reduce the symmetric matrix $\mathbf{A}\mathbf{A}^*$ to diagonal form. The diagonal elements of the resultant diagonal matrix being nonnegative, we can represent it as the square of another real diagonal matrix:

$$\mathbf{K}^{-1}(\mathbf{A}\mathbf{A}^*)\mathbf{K} = \Lambda^2 \qquad (15)$$

(In equality (15) the matrix $\mathbf{K}$ is orthogonal and the matrix Λ diagonal.) We can always arrange the diagonal elements μ_j of the matrix Λ in such a way that $\mu_1 \geqslant \mu_2 \geqslant \ldots \geqslant \mu_n$ $(\mu_n \geqslant 0)$, and

therefore we shall assume that the first k ($0 \leqslant k \leqslant n$) elements are positive and the other equal to zero. Now let us introduce the number (column) vectors $\mathbf{v}_j = \frac{1}{\mu_j}\mathbf{w}_j$ ($j = 1, 2, \ldots, k$) where the column vector $\mathbf{w}_j$ is the jth column of the matrix $\mathbf{A^*K}$. Since it follows from (15) that these vectors are of unit length and mutually orthogonal, we can add some number vectors $\mathbf{v}_{k+1}, \ldots, \mathbf{v}_n$ to them so as to form an orthogonal matrix $\mathbf{V}$; this can practically be realized by applying the orthogonalization process (e.g. see [107], Sec. VII.21). The construction of the matrix $\mathbf{V}$ implies that

$$\mathbf{A^*K} = \mathbf{V\Lambda} \text{ whence } \mathbf{A} = \mathbf{K\Lambda V^*} = \mathbf{K\Lambda L}$$

where $\mathbf{L} = \mathbf{V^*}$. The validity of representation (14) has thus been proved.

7. Commuting Mappings.

Lemma. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be square matrices of order* n. *Suppose that the matrix* $\mathbf{A}$ *is diagonal and that the first* n_1 *elements of its principal diagonal are equal to* λ_1, *the subsequent* n_2 *elements are equal to* λ_2, *etc., where all the numbers* λ_k *are pairwise distinct. Then for the equality* $\mathbf{AB} = \mathbf{BA}$ *to be fulfilled it is necessary and sufficient that* $\mathbf{B}$ *be a quasi-diagonal matrix with the orders* n_1, n_2, ... *of its submatrices arranged along the principal diagonal.*

Indeed, we readily verify the validity of the general equalities

$$\begin{pmatrix} \alpha_1 & 0 & \ldots & 0 \\ 0 & \alpha_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \alpha_n \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \ldots & b_{1n} \\ b_{21} & b_{22} & \ldots & b_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ b_{n1} & b_{n2} & \ldots & b_{nn} \end{pmatrix} = \begin{pmatrix} \alpha_1 b_{11} & \alpha_1 b_{12} & \ldots & \alpha_1 b_{1n} \\ \alpha_2 b_{21} & \alpha_2 b_{22} & \ldots & \alpha_2 b_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_n b_{n1} & \alpha_n b_{n2} & \ldots & \alpha_n b_{nn} \end{pmatrix}$$

and

$$\begin{pmatrix} b_{11} & b_{12} & \ldots & b_{1n} \\ b_{21} & b_{22} & \ldots & b_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ b_{n1} & b_{n2} & \ldots & b_{nn} \end{pmatrix} \begin{pmatrix} \alpha_1 & 0 & \ldots & 0 \\ 0 & \alpha_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \alpha_n \end{pmatrix} = \begin{pmatrix} \alpha_1 b_{11} & \alpha_2 b_{12} & \ldots & \alpha_n b_{1n} \\ \alpha_1 b_{21} & \alpha_2 b_{22} & \ldots & \alpha_n b_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_1 b_{n1} & \alpha_2 b_{n2} & \ldots & \alpha_n b_{nn} \end{pmatrix}$$

For the right members of these relations to coincide it is necessary and sufficient that the equalities $\alpha_j b_{jk} = \alpha_k b_{jk}$ be fulfilled for all j and k, which is equivalent to the requirement that for $\alpha_j \neq \alpha_k$ the number b_{jk} should be equal to zero. The latter fact implies the assertion of the lemma.

Now suppose that we are given an arbitrary set A, B, C, ... of linear mappings of a complex linear space (R) into itself, each of the mappings having Jordan submatrices only of the first order, i.e. having a diagonal matrix in an appropriately chosen basis.

Then *all these mappings are pairwise commuting* (i.e. the equalities $AB=BA$, $AC=CA$, $BC=CB$, ... take place) *if and only if there exists a common basis in which every mapping has a diagonal matrix.*

The "if" part of the proposition directly follows from the fact that the diagonal matrices are always commuting. To prove the "only if" part let us suppose that all the given mappings are pairwise commuting and choose a basis in (R) such that the matrix of the mapping A takes the form indicated in the lemma. Then it will follow from the lemma that the subspace (R_1) spanned by the first n_1 base vectors (it coincides with the subspace of all eigenvectors of the mapping A corresponding to the eigenvalue λ_1), the subspace (R_2) spanned by the subsequent n_2 base vectors, etc., are all invariant under each of the given mappings. Therefore, the mapping B can be considered separately in each subspace (R_j), and we can take a basis in (R_j) such that the matrix of the mapping B (which in (R_j) is of order n_j) takes the form described in the lemma. This transformation of the basis within each subspace (R_j) does not affect the form of the matrix of the mapping A (why?) and leads to the decomposition of each subspace (R_j) into the direct sum of subspaces (R_{jk}) invariant under every mapping. We then consider the mapping C and construct the corresponding basis in each subspace (R_{jk}) and so on. But these successive decompositions into direct sums cannot be performed more than $n-1$ times (why?), and hence, after a finite number of steps (even if there is an infinitude of the given mappings), we arrive at a basis in which the matrices of all the given mappings take diagonal form.

§ 4. NUMERICAL METHODS

There are a large number of numerical methods applied to solving various classes of problems of linear algebra which are treated in courses of computational methods (for instance, we can refer the reader to [9], [23], [77]). One of the most comprehensive books devoted to these questions is [36]. In § 4 we shall present a few of these methods. Naturally, when performing practical computations we pass from general mappings and spaces involved to the corresponding number vectors and matrices and operations on them.

The computational methods of linear algebra mainly deal with the problems of solving systems of linear equations and computing spectra, that is sets of eigenvalues of matrices.

1. Gauss' Method. The basic method of numerical solution of systems of linear equation is **Gauss' method (the elimination scheme**; e.g. see [107], Sec. VI.5).

Let us begin with a formal description of the simplest variant of Gauss' method applicable to systems of linear equations of the form

$$\left.\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{aligned}\right\} \tag{1}$$

which can briefly be written in the vector-matrix notation as

$$\mathbf{Ax} = \mathbf{b} \tag{2}$$

(the sense of the matrix $\mathbf{A}$ and number vectors $\mathbf{x}$ and $\mathbf{b}$ is obvious here). Dividing the first equation of system (1) by a_{11} we obtain the relation

$$x_1 + a'_{12}x_2 + \ldots + a'_{1n}x_n = b'_1 \tag{3}$$

This relation is then independently multiplied by $-a_{21}$, $-a_{31}$, ..., $-a_{n1}$ and, respectively, added to the second, third, ..., nth equations of system (1), which results in the equations of the form

$$\left.\begin{aligned} a'_{22}x_2 + \ldots + a'_{2n}x_n &= b'_2 \\ \cdots\cdots\cdots\cdots & \\ a'_{n2}x_2 + \ldots + a'_{nn}x_n &= b'_n \end{aligned}\right\} \tag{4}$$

Thus, the unknown x_1 has been eliminated. The next step is to divide the first equation of system (4) by a'_{22} which yields

$$x_2 + a''_{23}x_3 + \ldots + a''_{2n}x_n = b''_n$$

and to eliminate x_2, etc., until we arrive at a single equation in one unknown x_n. Then we find x_n from the last equation, x_{n-1} from the foregoing equation and so on, and finally, knowing the values of x_n, ..., x_2, we determine x_1 from equation (3), which completes the process of solution.

For the first step in the above scheme to be realizable it is necessary that the condition $a_{11} \neq 0$ be fulfilled. In system (4) we have

$$a'_{22} = a_{22} - a_{21}a'_{12} = a_{22} - a_{21}\frac{a_{12}}{a_{11}} = \frac{\Delta_2}{\Delta_1}$$

where Δ_k is the kth principal minor of the matrix $\mathbf{A}$ (see Sec. 2.3) and hence, for the second step of the above variant of Gauss' scheme to be realizable we must have $\Delta_2 \neq 0$. It can readily be shown that the subsequent coefficients by which we divide the equations are equal to $\frac{\Delta_3}{\Delta_2}$, $\frac{\Delta_4}{\Delta_3}$, etc. Consequently, to guarantee the possibility of completing the calculations in accordance with

this scheme we must impose the necessary and sufficient condition that all the principal minors of the matrix are different from zero. Furthermore, even if this condition holds but one of the principal minors is very close to zero the accuracy of the computations may be considerably diminished. This situation is readily detected in the computation process, and then we can divide the corresponding equation by some other coefficient which is not so small. This obviously is equivalent to renumbering the unknowns.

We can easily show that, from the point of view of matrix algebra, Gauss' scheme is equivalent to a representation of the (coefficient) matrix **A** of equation (2) in the form

$$\mathbf{A} = \mathbf{HB} \tag{5}$$

where **H** (**B**) is a **triangular (semidiagonal) matrix** whose all elements above (below) the principal diagonal are equal to zero. Here the rows of the matrix **B** are composed of the coefficients of equations (3), (4), etc., that is the elements of its principal diagonal are equal to unity. This fact is implied by the formulas (let the reader verify them!)

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} - \begin{pmatrix} k & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} a_{11}k^{-1} & a_{12}k^{-1} & \dots & a_{1n}k^{-1} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 0 & \dots & 0 \\ -\alpha & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ -\gamma & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21}+\alpha a_{11} & a_{22}+\alpha a_{12} & \dots & a_{2n}+\alpha a_{1n} \\ \dots & \dots & \dots & \dots \\ a_{n1}+\gamma a_{11} & a_{n2}+\gamma a_{12} & \dots & a_{nn}+\gamma a_{1n} \end{pmatrix}$$

which indicate that every step in Gauss' method is equivalent to representing the corresponding coefficient matrix in the form of a product of two matrices the left of which is a triangular matrix of the type of **H**. A product of such semidiagonal matrices is again a matrix of that type (why?), and hence, after all the steps have been made, we arrive at representation (5).

When system (2) has been reduced to the form $\mathbf{HBx} = \mathbf{b}$ it can be readily solved if we put $\mathbf{Bx} = \mathbf{y}$ and thus pass to the two consecutive **triangular** systems $\mathbf{Hy} = \mathbf{b}$ and $\mathbf{Bx} = \mathbf{y}$. By the way, if we apply a modification of Gauss' method in which certain elimination operations are performed not only on the left-hand but also on the right-hand sides of system (2) we can directly obtain the triangular system $\mathbf{Bx} = \mathbf{H}^{-1}\mathbf{b}$ whose right-hand side is known.

Some useful considerations concerning Gauss' method can be found in [38].

2. Norm of a Matrix. Let a linear mapping A of a real Euclidean space (R) into itself have a (square) matrix **A** in a Cartesian basis consisting of mutually orthogonal unit vectors. Then the **norm $\|A\|$ of the mapping** A (also spoken of as the **norm $\|A\|$ of the matrix A**) is defined as the greatest magnification coefficient corresponding to the stretching of the vectors under this mapping:

$$\|A\| = \|\mathbf{A}\| = \max_{x \in R} \frac{|Ax|}{|x|} = \max_{x \neq 0} \left\{ \left[\sum_j \left(\sum_k a_{jk} x_k \right)^2 \right]^{\frac{1}{2}} \left[\sum_j x_j^2 \right]^{-\frac{1}{2}} \right\} \quad (6)$$

We can readily establish the following properties of the norm of a matrix (prove them!):

1. $\|\boldsymbol{O}\| = 0$, and the norms of the other matrices are positive.
2. For any $\boldsymbol{x} \in (R)$ we have $|\mathbf{A}\boldsymbol{x}| \leqslant \|\mathbf{A}\| \, |\boldsymbol{x}|$, the vectors $\boldsymbol{x} \neq \boldsymbol{0}$ for which this inequality turns into the equality always existing.
3. $\|\mathbf{A}+\mathbf{B}\| \leqslant \|\mathbf{A}\| + \|\mathbf{B}\|$.
4. $\|\lambda\mathbf{A}\| = |\lambda| \, \|\mathbf{A}\|$.
5. $\|\mathbf{AB}\| \leqslant \|\mathbf{A}\| \cdot \|\mathbf{B}\|$, which, in particular, implies that $\|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \geqslant \|\mathbf{I}\| = 1$.
6. The norm of every orthogonal matrix is equal to 1.
7. The norm of a symmetric matrix is equal to the greatest absolute value of its eigenvalues.
8. The square of the norm of any matrix **A** is equal to the greatest eigenvalue of the matrix $\mathbf{AA}^*$.

(When proving Properties 7 and 8 use the results of Sec. 3.6.)

These properties are readily reformulated in terms of the corresponding mappings and also extended to mappings of complex spaces and complex matrices. To find an upper bound for the norm we can take advantage of the following inequality implied by definition (6) and the Cauchy-Bunyakovsky-Schwarz inequality (e. g. see [107], relation (VII.26)):

$$\left(\sum_k a_{jk} x_k \right)^2 \leqslant \left(\sum_k a_{jk}^2 \right) \left(\sum_k x_k^2 \right)$$

From this inequality we obtain

$$\|\mathbf{A}\| \leqslant \left[\sum_j \left(\sum_k a_{jk}^2 \right) \right]^{\frac{1}{2}}$$

A lower bound for the norm can be estimated by using the consequence of Property 5 (see above).

The norm of a matrix is in particular used for introducing the notion of a *well-conditioned linear system* of type (2) which is important for computations. Suppose that the right-hand side of (2) is known with an error $\Delta\mathbf{b}$; then the solution is also determined to within an error $\Delta\mathbf{x}$, and $\mathbf{A}(\mathbf{x}+\Delta\mathbf{x}) = \mathbf{b}+\Delta\mathbf{b}$ (because, for simplicity, the coefficients of the system are considered

as being absolutely precise). Consequently, we have $\mathbf{A}\Delta\mathbf{x}=\Delta\mathbf{b}$, which, together with (2), implies

$$\Delta\mathbf{x}=\mathbf{A}^{-1}\Delta\mathbf{b},\ |\Delta\mathbf{x}|\leqslant\|\mathbf{A}^{-1}\|\,|\Delta\mathbf{b}|,\ |\mathbf{b}|\leqslant\|\mathbf{A}\|\,|\mathbf{x}| \tag{7}$$

whence

$$\frac{|\Delta\mathbf{x}|}{|\mathbf{x}|}\leqslant\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|\,\frac{|\Delta\mathbf{b}|}{|\mathbf{b}|} \tag{8}$$

where both inequalities (7), and therefore inequality (8) as well, turn into equalities for some nonzero values of $\mathbf{b}$ and $\Delta\mathbf{b}$.

We see that if the product $\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$ is close to unity (it obviously cannot be less than unity) the solutions of (2) are stable with respect to errors in its right-hand side because this quantity is equal to the greatest possible magnification coefficient of the relative error of the solution in comparison with the relative error of the column vector formed of the right-hand sides. In this case we say that system (2) and the matrix $\mathbf{A}$ are **well-conditioned.** The quantity $\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$ serves as measure of this property and is called the **condition number** of the matrix $\mathbf{A}$. For a well-conditioned system the relative error of the solution is close to that of the known parameters of the system. If the condition number $\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$ is large the solution of system (2) is unstable in the sense that the relative error in determining the solution may considerably exceed that of the parameters of the system. Therefore these systems are said to be **ill-conditioned** and are inconvenient for numerical computations.

It is readily seen (prove it!) that the condition member of a matrix $\mathbf{A}$ is equal to the ratio of the greatest semiaxis of the ellipsoid which is the image of a sphere under the mapping A to its smallest semiaxis. The square of the condition number is equal to the ratio of the greatest eigenvalue of the matrix $\mathbf{AA}^*$ to its smallest eigenvalue.

We thus conclude that the accuracy achieved in solving a linear system is specified by the condition number $\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$ and not purely by the magnitude of the determinant of the system. It is sometimes said that the accuracy decreases when the determinant of the system is small (e.g. see [107], Sec. VI.6), but this formulation is not precise because we can always multiply all the equations of the given system by a large common factor and thus make its determinant as large as desired (without affecting the magnitude of $\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$). The more precise formulation given above indicates that the accuracy increases when the system is well-conditioned and may decrease when the condition number $\|\mathbf{A}\|\,\|\mathbf{A}\|^{-1}$ is large.

It should also be noted that the condition number of the coefficient matrix may be affected when the system in question is subjected to some simple transformations, for instance, when one of

the equations is multiplied by a number while the other equations remain unchanged, and the like. For example, let the reader verify that for the system

$$kx+2ky=k,\ x+y=2$$

we have $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|=\frac{5k^2+2+\sqrt{25k^4+16k^2+4}}{2|k|}$ $\left(\mathbf{A}=\begin{pmatrix}k & 2k\\ 1 & 1\end{pmatrix}\right)$ although the solution is independent of k. Find the value of k for which the coefficient matrix of this system is **best conditioned** (i.e. the condition number $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ assumes its least value) and explain the fact that $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|\longrightarrow\infty$ for $k\longrightarrow 0$ and for $k\longrightarrow\pm\infty$.

In conclusion we note that there are some other definitions of the norm of a matrix and of the condition number which are used in applications.

3. Method of Minimizing Discrepancy. The round-off errors may result in an approximate solution $\mathbf{x}_1$ (of system (1)) whose accuracy is insufficient even when we apply the "precise" Gauss method. Putting $\mathbf{x}=\mathbf{x}_1+\mathbf{z}$ we obtain for the error $\mathbf{z}$ the equation

$$\mathbf{Az}=\mathbf{A}(\mathbf{x}-\mathbf{x}_1)=\mathbf{b}-\mathbf{Ax}_1,\quad \text{i.e. } \mathbf{Az}=\mathbf{r}_1 \tag{9}$$

where $\mathbf{r}_1=\mathbf{b}-\mathbf{Ax}_1$ is the discrepancy corresponding to the approximate solution $\mathbf{x}_1$. If $|\mathbf{r}_1|\ll|\mathbf{b}|$ (which is most often the case) equation (9) can be solved with a considerably greater absolute precision than (2). This makes it possible to increase the relative precision of the original approximate solution (why?).

It is clear that the problem of solving equation (2) is equivalent to that of determining the minimum of the square of the discrepancy, that is the minimum of the function

$$F(x_1, x_2, \ldots, x_n)=|\mathbf{Ax}-\mathbf{b}|^2 \tag{10}$$

(the square is taken for the sake of convenience of computations), and therefore we can apply to this problem direct minimization methods. In particular, we can use the *method of steepest descent* (e.g. see [107], Sec. XII.10). We have $\operatorname{grad}(\mathbf{x}, \mathbf{a})=\mathbf{a}$, and consequently

$$\operatorname{grad}|\mathbf{Ax}-\mathbf{b}|^2=2\operatorname{grad}(\mathbf{Ax}-\mathbf{b},\ (\mathbf{Ax}-\mathbf{b})_{\text{const}})=$$
$$=2\operatorname{grad}(\mathbf{x},\ \mathbf{A}^*(\mathbf{Ax}-\mathbf{b})_{\text{const}})=2\mathbf{A}^*(\mathbf{Ax}-\mathbf{b})$$

where the subscript "const" indicates the quantities which are not differentiated when the operation of taking the gradient is applied. Hence, every subsequent approximation in this method is expressed in terms of the preceding one by the formula

$$\mathbf{x}_{k+1}=\mathbf{x}_k-2t\mathbf{A}^*(\mathbf{Ax}_k-\mathbf{b})=\mathbf{x}_k+2t\mathbf{A}^*\mathbf{r}_k$$

where $t=t_k$ is chosen in accord with the condition that function

(10) is minimized. Applying the necessary condition for minimum with respect to the variable t (the calculations are left to the reader) we arrive at the formula

$$\mathbf{x}_{k+1} = \mathbf{x}_k + |\mathbf{A}^*\mathbf{r}_k|^2 |\mathbf{A}\mathbf{A}^*\mathbf{r}_k|^{-2} \mathbf{A}^*\mathbf{r}_k,$$

which enables us to perform iterations beginning with an arbitrary zero (initial) approximation.

In the particular case when the matrix $\mathbf{A}$ is symmetric and the quadratic form $(\mathbf{Ax}, \mathbf{x})$ is positive definite the solution of equation (2) coincides with the only minimum of the function $(\mathbf{Ax}-2\mathbf{b}, \mathbf{x})$. Prove this assertion and construct, on this basis, a more convenient iterative scheme for solving equation (2) for which in this case we have

$$\mathbf{x}_{k+1} = \mathbf{x}_k + |\mathbf{r}_k|^2 (\mathbf{A}\mathbf{r}_k, \mathbf{r}_k)^{-1} \mathbf{r}_k$$

In particular, matrices of that type are encountered in the *calculus of variations* (see Sec. VI.4.1).

4. Spectrum of a Symmetric Matrix. Now we proceed to consider the problem of finding the spectrum (the set of the eigenvalues) of a given matrix $\mathbf{A}$ and begin with the case when $\mathbf{A}$ is a real symmetric matrix. Its greatest and least eigenvalues (these play the most important role in many applied problems concerning the spectrum) can be determined numerically with the aid of the extremal property given in Sec. 1.7. Let us demonstrate the application of the method of steepest descent to this scheme. To this end, let us arbitrarily choose an initial vector $\mathbf{x}_0$ of unit length. The hodograph of the vector function

$$\mathbf{x} = \frac{\mathbf{x}_0 + \operatorname{grad}(\mathbf{Ax}, \mathbf{x})|_{\mathbf{x}=\mathbf{x}_0} t}{|\mathbf{x}_0 + \operatorname{grad}(\mathbf{Ax}, \mathbf{x})|_{\mathbf{x}=\mathbf{x}_0} t|} = \frac{\mathbf{x}_0 + 2\mathbf{A}\mathbf{x}_0 t}{|\mathbf{x}_0 + 2\mathbf{A}\mathbf{x}_0 t|} \tag{11}$$

is an arc of a great circle of unit sphere, and for $t > 0$ the motion along this arc is in the direction of the fastest growth of the function $(\mathbf{Ax}, \mathbf{x})$ on that sphere. Let us take a value of t such that the quantity

$$(\mathbf{Ax}, \mathbf{x}) = \frac{(\mathbf{A}(\mathbf{x}_0 + 2\mathbf{A}\mathbf{x}_0 t), \mathbf{x}_0 + 2\mathbf{A}\mathbf{x}_0 t)}{(\mathbf{x}_0 + 2\mathbf{A}\mathbf{x}_0 t, \mathbf{x}_0 + 2\mathbf{A}\mathbf{x}_0 t)} \tag{12}$$

takes on its greatest or least value. This is equivalent to finding conditional extrema of the function $(\mathbf{Ax}, \mathbf{x})$ on that arc. The application of the necessary condition for conditional extremum to (12) as function of t (the calculations are left to the reader) results in the equation

$$4(\alpha_0\gamma_0 - \beta_0^2) t^2 + 2(\alpha_0\delta_0 - \beta_0\gamma_0) t + (\beta_0\delta_0 - \gamma_0^2) = 0 \tag{13}$$

in which

$$\alpha_0 = (\mathbf{A}^3\mathbf{x}_0, \mathbf{x}_0), \quad \beta_0 = (\mathbf{A}^2\mathbf{x}_0, \mathbf{x}_0), \quad \gamma_0 = (\mathbf{A}\mathbf{x}_0, \mathbf{x}_0), \quad \delta_0 = (\mathbf{x}_0, \mathbf{x}_0)$$

The substitution of the positive and negative roots of equation (13) into the right-hand side of (11) yields, respectively, the points of relative maximum and minimum. Then the procedure described above is repeatedly applied to the vectors thus obtained, etc. (in this application we must again take the direction in which the variation of the quantity $(\mathbf{Ax}, \mathbf{x})$ is of the desired type). After a sufficient number of iterations have been performed, we not only obtain the desired approximations to the greatest and the least eigenvalues of the matrix $\mathbf{A}$ but also to the corresponding eigenvectors.

If we are also interested in the intermediate eigenvalues and the corresponding eigenvectors we can add to the two eigenvectors $\mathbf{l}_1$ and $\mathbf{l}_2$ thus found some vectors $\mathbf{l}_3, \mathbf{l}_4, \ldots$ such that the system $\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_n$ is a Euclidean basis, and then express the quadratic form $(\mathbf{Ax}, \mathbf{x})$ in the new coordinates: $(\mathbf{Ax}, \mathbf{x}) = (\mathbf{A'y}, \mathbf{y})$. Since the eigenvectors are mutually orthogonal, we put $y_1 = y_2 = 0$ and thus pass to the corresponding quadratic form in $n-2$ coordinates and again apply the above procedure to it, etc.

In some problems it is required to find the least absolute value of the eigenvalues of a matrix $\mathbf{A}$. Then we can determine the least eigenvalue of the matrix $\mathbf{A}^2$ and extract the square root of it (why?).

Some useful qualitative and quantitative considerations related to the problem of determining spectra of matrices can be found in [71].

5. Method of K.G.J. Jacobi. Here we shall present an iterative method for computing the eigenvalues and eigenvectors of a real symmetric matrix $\mathbf{A}$ inaugurated by K.G.J. Jacobi in 1846. The geometric idea of the method can be explained as follows. As was shown in Sec. 2.1, the equation $\mathbf{x^*Ax} = 1$ determines, in the space E_n, a surface (M) of the second order. If, for definiteness, we suppose that this surface is an ellipsoid, the problem reduces to determining the principal axes of the ellipsoid. Let $\mathbf{l}_1, \ldots, \mathbf{l}_n$ be an initial basis chosen in E_n such that the pth coordinate of $\mathbf{l}_p (p = 1, 2, \ldots, n)$ is equal to unity while the other coordinates are equal to zero. Choose two arbitrary base vectors $\mathbf{l}_i$ and $\mathbf{l}_j$ $(i \neq j)$ of this basis and consider the plane spanned by these vectors. This plane intersects (M) along an ellipse $(M)_{ij}$. Now let us substitute the unit vectors $\mathbf{l}'_i$ and $\mathbf{l}'_j$ directed along the principal axes of the ellipse $(M)_{ij}$ for the vectors $\mathbf{l}_i$ and $\mathbf{l}_j$ without changing the other base vectors. Then in the new basis thus obtained (we denote it as $\mathbf{l}'_1, \ldots, \mathbf{l}'_n$) we take two arbitrary base vectors and again turn them, in the plane of these vectors, in such a way that they go along the principal axes of the ellipse obtained in the section of (M) by that plane, etc.

Generally speaking, this process may continue indefinitely. Indeed, suppose, for instance, that $n = 3$ and the vectors $\mathbf{l}'_1$ and $\mathbf{l}'_2$

are obtained from $\mathbf{l}_1$ and $\mathbf{l}_2$ by rotating the latter about $\mathbf{l}_3$ as axis, and $\mathbf{l}_3' = \mathbf{l}_3$. Furthermore, let $\mathbf{l}_2''$ and $\mathbf{l}_3''$ be obtained from $\mathbf{l}_2'$ and $\mathbf{l}_3'$ by turning them about $\mathbf{l}_1'$, and $\mathbf{l}_1'' = \mathbf{l}_1'$. But these vectors $\mathbf{l}_1''$ and $\mathbf{l}_2''$ are no longer directed along the principal axes of the ellipse obtained in the plane of these vectors, and therefore we must again make a rotation, etc. The principal axes of the ellipsoid are obtained only in the limit.

This method can also be described analytically. We choose an element $a_{ij} \neq 0$ $(i \neq j)$ of the matrix $\mathbf{A}$ (it is preferable that a_{ij} belong to those elements whose absolute values are comparatively greater). Putting all the coordinates except x_i and x_j equal to zero we reduce the function $\mathbf{x}^*\mathbf{A}\mathbf{x}$ to the form

$$a_{ii}x_i^2 + 2a_{ij}x_ix_j + a_{jj}x_j^2$$

As is known from analytic geometry (e.g. see [107], Sec. II.13), to reduce this expression to diagonal form it is necessary to turn the coordinate axes through the angle φ_1 determined by the relation $\tan 2\varphi_1 = \frac{2a_{ij}}{a_{ii} - a_{jj}}$. The corresponding transformation of the coordinates is written as $x_i = x_i' \cos\varphi_1 - x_j' \sin\varphi_1$, $x_j = x_i' \sin\varphi_1 + x_j' \cos\varphi_1$, $x_s = x_s'$ for $s \neq i, j$. In the new coordinates the matrix of the quadratic form is equal to $\mathbf{A}' = \mathbf{H}_1^*\mathbf{A}\mathbf{H}_1$ where the elements h_{ij} of the matrix $\mathbf{H}_1$ are the following: $h_{ii} = h_{jj} = \cos\varphi_1$, all the other diagonal elements are equal to unity; $h_{ij} = -\sin\varphi_1$, $h_{ji} = \sin\varphi_1$ and all the other off-diagonal elements are equal to zero. Now we choose an element $a_{rk}' \neq 0$ $(r \neq k)$ of the matrix $\mathbf{A}'$ and perform the rotation of the $x_r'x_k'$-plane through the angle φ_2 determined by the equality $\tan 2\varphi_2 = \frac{2a_{rk}'}{a_{rr}' - a_{kk}'}$, which reduces the matrix to the form $\mathbf{A}'' = \mathbf{H}_2^*\mathbf{A}'\mathbf{H}_2$ where the structure of the matrix $\mathbf{H}_2$ is analogous to that of $\mathbf{H}_1$, etc. We proceed in this way until the nondiagonal elements of the matrix $\mathbf{A}^{(m)}$ have become sufficiently small in their absolute values. Then the iteration process is stopped and the diagonal elements of the matrix $\mathbf{A}^{(m)}$ are taken as approximate eigenvalues of the matrix $\mathbf{A}$. Besides, since $\mathbf{A}^{(m)} = -\mathbf{H}^*\mathbf{A}\mathbf{H}$ where $\mathbf{H} = \mathbf{H}_1\mathbf{H}_2 \ldots \mathbf{H}_m$, the original coordinates are connected with the terminal ones by the relation $\mathbf{x} = \mathbf{H}\mathbf{x}^{(m)}$, and hence the approximate eigenvectors of the matrix $\mathbf{A}$ corresponding to the eigenvalues thus found coincide with the columns of the matrix $\mathbf{H}$ (why?).

6. Iteration Method for Computing Eigenvalues with the Greatest Modulus. Now we pass to general square matrices whose elements may be complex. Suppose we are given a matrix $\mathbf{A}$ whose all eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ (which are unknown) are simple, the absolute value $|\lambda_1|$ being (strictly) greater than the other $|\lambda_k|$'s.

Let the corresponding eigenvectors (which are also unknown) be denoted by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Choose an arbitrary vector $\mathbf{y}_0 \neq \mathbf{0}$ and consider the sequence of vectors $\{\mathbf{y}_m\}$ constructed in accordance with the formulas $\mathbf{y}_1 = \mathbf{A}\mathbf{y}_0$, $\mathbf{y}_2 = \mathbf{A}\mathbf{y}_1$, ... The vector $\mathbf{y}_0$ can be resolved in the form $\mathbf{y}_0 = \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \ldots + \alpha_n\mathbf{x}_n$, and consequently we have

$$\mathbf{y}_1 = \mathbf{A}\mathbf{y}_0 = \alpha_1\mathbf{A}\mathbf{x}_1 + \alpha_2\mathbf{A}\mathbf{x}_2 + \ldots + \alpha_n\mathbf{A}\mathbf{x}_n = \\ = \alpha_1\lambda_1\mathbf{x}_1 + \alpha_2\lambda_2\mathbf{x}_2 + \ldots + \alpha_n\lambda_n\mathbf{x}_n,$$
$$\mathbf{y}_2 = \mathbf{A}\mathbf{y}_1 = \alpha_1\lambda_1^2\mathbf{x}_1 + \alpha_2\lambda_2^2\mathbf{x}_2 + \ldots + \alpha_n\lambda_n^2\mathbf{x}_n$$

and, generally,

$$\mathbf{y}_m = \alpha_1\lambda_1^m\mathbf{x}_1 + \alpha_2\lambda_2^m\mathbf{x}_2 + \ldots + \alpha_n\lambda_n^m\mathbf{x}_n = \\ = \lambda_1^m\left[\alpha_1\mathbf{x}_1 + \alpha_2\left(\frac{\lambda_2}{\lambda_1}\right)^m\mathbf{x}_2 + \ldots + \alpha_n\left(\frac{\lambda_n}{\lambda_1}\right)^m\mathbf{x}_n\right]$$

By the hypothesis concerning $|\lambda_1|$, we have, for large m, the asymptotic relation $\mathbf{y}_m \sim \lambda_1^m\alpha_1\mathbf{x}_1$. Hence, if we denote by $y_m^{(k)}$ the kth element of the column vector $\mathbf{y}_m$, the value of λ_1 is expressed by the formula

$$\lambda_1 = \lim_{m\to\infty} \frac{y_{m+1}^{(k)}}{y_m^{(k)}} \tag{14}$$

provided that $\alpha_1 x_1^{(k)} \neq 0$. As the eigenvector $\mathbf{x}_1$ we then can take the limit $\lim_{m\to\infty}(\mu_m\mathbf{y}_m)$ where μ_m are normalization factors (μ_m can be taken equal to $\frac{1}{|\mathbf{y}_m|}$ or to $\frac{1}{|y_m^{(k)}|}$, etc.).

Considering the associated vectors (entering into the construction of the corresponding Jordan canonical matrix) we can show that for this method to be realizable it is only essential that the above condition on $|\lambda_1|$ and the requirement that λ_1 is a simple eigenvalue are fulfilled (more precisely, the Jordan submatrices corresponding to λ_1 must not be of order higher than the first) while the other eigenvalues can be multiple. As to the fulfilment of these essential assumptions, it is indicated in the computation process by the mere fact that limit (14) exists; this limit must be dependent neither on k (except some special cases when $x_1^{(k)} = 0$ which are indicated by a slower increase of $y_m^{(k)}$) nor on $\mathbf{y}_0$ (except some rarely encountered cases when $\alpha_1 = 0$; if the elements of $\mathbf{y}_0$ are taken at random we can be sure that this will not occur).

If there are two distinct simple eigenvalues λ_1 and λ_2 with the same greatest modulus, and $\lambda_2 = -\lambda_1$, we obtain, for large values of m, the relation $\mathbf{y}_m \sim \lambda_1^m[\alpha_1\mathbf{x}_1 + (-1)^m\alpha_2\mathbf{x}_2]$, that is there are two different limits of form (14) for odd and even values of m; it is this fact that makes it possible to recognize the situation with

two such eigenvalues λ_1 and λ_2. Here we can put

$$\lambda_1^2 = \lim_{m \to \infty} \frac{y_{m+2}^{(k)}}{y_m^{(k)}}$$

Up till now we have dealt with the general case of a complex matrix although for real matrices the calculations are of course considerably simpler. For a real matrix **A** there is another important case when there appear two simple imaginary conjugate complex eigenvalues with the greatest modulus, i.e. $\lambda_2 = \lambda_1^* \neq \lambda_1$. Then the corresponding elements of the eigenvectors associated with these eigenvalues can be chosen so that they are the complex conjugates of each other, and then for any real $\mathbf{y}_0$ we have, for large values of m, the relation

$$y_m^{(k)} \sim \lambda_1^m \alpha_1 x_1^{(k)} + \lambda_1^{*m} \alpha_1^* x_1^{(k)*} = 2Ar^m \cos(ma + b) \tag{15}$$

where $\alpha_1 x_1^{(k)} = Ae^{ib}$ and $\lambda_1 = re^{ia}$. In this case the sign of $y_m^{(k)}$ varies in a complicated manner when m increases, and this fact makes it possible to recognize a situation of that kind. If we write the quadratic equation

$$\lambda^2 + p\lambda + q = 0 \tag{16}$$

with real coefficients whose roots are λ_1 and λ_1^* and replace the symbol $\sim$ in relations (15) by the sign of equality it is readily seen (prove it!) that

$$y_{m+2}^{(k)} + py_{m+1}^{(k)} + qy_m^{(k)} = 0$$

Making k in the latter relation assume two different values we can determine the values of p and q which must tend, for $m \to \infty$, to their limits independent of the choice of k; after that we find from (16) the values of λ_1 and λ_1^*. The corresponding eigenvector $\lambda_1^m \alpha_1 \mathbf{x}_1$ is then found from (15) after we substitute for m two subsequent large values.

7. Computing Subsequent Eigenvalues. As is shown below, if an eigenvector $\mathbf{x}_1$ belonging to an eigenvalue of an $(n \times n)$-matrix **A** is known, the problem of determining the other eigenvectors and eigenvalues can be reduced to the similar problem for a matrix $\mathbf{A}_1$ of order $n-1$. Since the method of Sec. 6 makes it possible to find the eigenvalue of **A** with the greatest modulus, we can in fact pass to the corresponding matrix $\mathbf{A}_1$, then again reduce the order by unity and so on.

To pass from **A** to the matrix $\mathbf{A}_1$, let us denote the first row of the matrix **A** by **r** (and regard it as a row vector) and also normalize the eigenvectors $\mathbf{x}_k$ of **A** (which are defined to within nonzero scalar factors) in such a way that their first coordinates are equal to unity. Let us put $\mathbf{B} = \mathbf{A} - \mathbf{x}_1\mathbf{r}$ (here and below the n-dimensional row and column vectors are, respectively, understood

as $(1\times n)$- and $(n\times 1)$-matrices, and their multiplication is performed according to the rules of matrix multiplication; e.g. see [107], Sec. XI.1.2). The above normalization condition implies that the first row of the matrix $\mathbf{B}$ consists of zeros (why?). It is also obvious that $\mathbf{r}\mathbf{x}_k=\lambda_k$ (let the reader prove this equality). It follows from this equality and the relation

$$\mathbf{A}\mathbf{x}_k=\lambda_k\mathbf{x}_k$$

that

$$\mathbf{B}(\mathbf{x}_k-\mathbf{x}_1)=\mathbf{A}\mathbf{x}_k-\mathbf{x}_1\mathbf{r}\mathbf{x}_k-\mathbf{A}\mathbf{x}_1+\mathbf{x}_1\mathbf{r}\mathbf{x}_1=$$
$$=\lambda_k\mathbf{x}_k-\lambda_k\mathbf{x}_1-\lambda_1\mathbf{x}_1+\lambda_1\mathbf{x}_1=\lambda_k(\mathbf{x}_k-\mathbf{x}_1)$$

that is

$$\mathbf{B}(\mathbf{x}_k-\mathbf{x}_1)=\lambda_k(\mathbf{x}_k-\mathbf{x}_1) \tag{17}$$

Let us denote by $\mathbf{y}_k$ the column of $n-1$ numbers obtained from the column vector $\mathbf{x}_k-\mathbf{x}_1$ by discarding its upper element (which is equal to zero). Now we choose as $\mathbf{A}_1$ the matrix of order $n-1$ obtained from $\mathbf{B}$ by deleting its first row and first column. Then (17) implies that $\mathbf{A}_1\mathbf{y}_k=\lambda_k\mathbf{y}_k$ (why?), that is the vector $\mathbf{y}_k$ ($k=2$, 3, ..., n) is an eigenvector of the matrix $\mathbf{A}_1$ belonging to the eigenvalue λ_k ($\mathbf{y}_1$ is a zero vector and is therefore of no interest). After the vector $\mathbf{y}_k$ has been found, we construct a column vector of n elements by introducing into $\mathbf{y}_k$ an additional (upper) element equal to zero. To the resultant vector we add $\mathbf{x}_1$ and thus obtain $\mathbf{x}_k$; in this construction the vector $\mathbf{y}_k$ is normalized by the condition $\mathbf{r}\mathbf{x}_k=\lambda_k$, and then relation (17) obviously implies that $\mathbf{A}\mathbf{x}_k=\lambda_k\mathbf{x}_k$.

It may turn out that the latter normalization is impossible. Then, for any normalization factor in $\mathbf{y}_k$, we have $\mathbf{r}(\mathbf{x}_k-\mathbf{x}_1)=0$, i.e. $\mathbf{r}\mathbf{x}_k=\lambda_1$. It can be readily shown that in this case the vector $\mathbf{x}_k-\mathbf{x}_1$ whose upper element is equal to zero serves as an eigenvector corresponding to the eigenvalue λ_k. This zero component accounts for the fact that the normalization is impossible.

If the upper element of $\mathbf{x}_1$ is equal to zero we can use, in this construction, any other element instead of the upper one. Therefore, the above method can always be applied when the eigenvalues of the matrix $\mathbf{A}$ are simple.

If $\mathbf{A}$ is a real matrix and its two complex conjugate eigenvalues and the pair of the corresponding eigenvectors $\mathbf{x}_{1,2}=\tilde{\mathbf{x}}_1\pm i\tilde{\mathbf{x}}_2$ have been found, it can be shown that the problem of determining the other eigenvalues and eigenvectors can be immediately reduced to the analogous problem for a matrix $\mathbf{A}_2$ of the $(n-2)$th order. For this purpose we normalize the vectors $\mathbf{x}_{1,2}$ by the condition $\tilde{a}_1\tilde{b}_2-\tilde{b}_1\tilde{a}_2=1$ where $\tilde{a}_j$ and $\tilde{b}_j$ are the first two (from the top) elements of the vector $\tilde{\mathbf{x}}_j$ ($j=1$, 2). Then the matrix $\mathbf{C}=\mathbf{A}-$

$-(\tilde{b}_2\tilde{\mathbf{x}}_1-\tilde{b}_1\tilde{\mathbf{x}}_2)\,\mathbf{r}+(\tilde{a}_2\tilde{\mathbf{x}}_1-\tilde{a}_1\tilde{\mathbf{x}}_2)\,\mathbf{s}$, where $\mathbf{r}$ and $\mathbf{s}$ are the first two rows of the matrix $\mathbf{A}$, has the first two rows composed of zeros. Deleting in $\mathbf{C}$ these rows and also the first two columns we obtain $\mathbf{A}_2$. If $\mathbf{z}_k$ is an eigenvector of $\mathbf{A}_2$ corresponding to the eigenvalue λ_k (i.e. $\mathbf{A}_2\mathbf{z}_k=\lambda_k\mathbf{z}_k$) we introduce into $\mathbf{z}_k$ two zero upper elements, add to the resulting vector a vector of the form $(u\tilde{b}_2-v\tilde{a}_2)\,\tilde{\mathbf{x}}_1+(v\tilde{a}_1-u\tilde{b}_1)\,\tilde{\mathbf{x}}_2$ and obtain an eigenvector $\mathbf{x}_k$ of the matrix $\mathbf{A}$ belonging to the eigenvalue λ_k if the scalars u and v are chosen in such a way that $\mathbf{r}\mathbf{x}_k=u\lambda_k$ and $\mathbf{s}\mathbf{x}_k=v\lambda_k$. The justification of this procedure is left to the reader.

8. **Matrices with Nonnegative Elements.** Matrices with real nonnegative elements are encountered in a number of applied problems. For these matrices a special theory has been developed which, in particular, makes it possible to obtain estimations for their greatest eigenvalues (e.g. see [44], [67]).

Let us take a matrix $\mathbf{A}$ satisfying the condition

$$a_{jk}\geqslant 0 \qquad (j,\ k=1,\ 2,\ \ldots,\ n) \tag{18}$$

Here we shall introduce, for vectors, an *order relation*. Namely, if $\mathbf{x}$ and $\mathbf{y}$ are vectors whose components satisfy the inequalities $x_1\geqslant y_1$, $x_2\geqslant y_2$, ..., $x_n\geqslant y_n$ we shall write $\mathbf{x}\succcurlyeq\mathbf{y}$ (read "the vector $\mathbf{y}$ precedes the vector $\mathbf{x}$ or coincides with it"). Note that, for the matrices $\mathbf{A}$ of this class, the relation $\mathbf{x}\succcurlyeq\mathbf{y}$ implies $\mathbf{A}\mathbf{x}\succcurlyeq\mathbf{A}\mathbf{y}$.

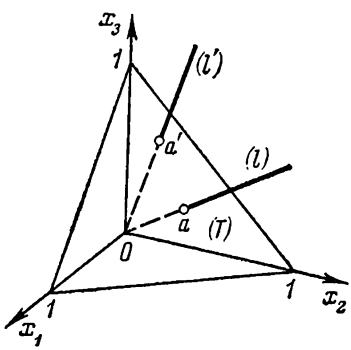


Fig. 64

We shall begin with proving the following assertion: *if* $\mathbf{A}$ *is a matrix satisfying condition* (18) *it possesses at least one eigenvector* $\mathbf{x}_0\succcurlyeq\mathbf{0}$; the eigenvalue corresponding to this eigenvector must be real and nonnegative (why?).

To simplify the proof let us assume that $n=3$ and regard $\mathbf{A}$ as the matrix of a mapping of a real three-dimensional Euclidean space (which can be interpreted as the three-dimensional geometric space; see Fig. 64) into itself. Then condition (18) means that every ray (l) with vertex at the origin O which lies in the first octant is transformed into a ray (l') in the same octant provided that, under the mapping, it does not go as a whole into the origin (why?). If the latter is the case the vectors parallel to (l) are eigenvectors (they correspond to the eigenvalue $\lambda=0$), and thus for this case our assertion is true. Therefore in what follows we exclude this variant.

We must prove that at least one ray in the first octant is mapped onto itself. To this end, let us consider the triangle (T) (see Fig. 64) determined by the conditions

$$x_1+x_2+x_3=1, \quad x_1 \geqslant 0, \quad x_2 \geqslant 0, \quad x_3 \geqslant 0$$

There is an obvious one-to-one correspondence between the rays in the first octant and the points of the triangle, and consequently every mapping of the rays into themselves induces a mapping of the triangle (T) into itself, the latter mapping being continuous. Now we can make use of the **fixed-point theorem** proved in 1904 by the Latvian mathematician P. Bohl (1865-1921) and independently in 1911-1913 by the Dutch mathematician L.E.J. Brouwer (1882-1966). The **Bohl-Brouwer theorem** states that *for any continuous mapping of a bounded convex finite-dimensional figure into itself there exists at least one* **fixed point** *of the mapping*, i.e. one that goes into itself (by the way, this also follows from the results established in Sec. VIII.2.6). This theorem implies that our assertion concerning the rays is true, and therefore the above-mentioned eigenvector $\mathbf{x}_0$ exists.

Let us now additionally suppose that inequalities (18) are strict (although the considerations given below also apply when non-strict inequalities (18) hold and all the elements of one of the matrices $\mathbf{A}^2$, $\mathbf{A}^3$, ... are nonzero). Then all the elements of the eigenvector $\mathbf{x}_0$ constructed in the foregoing paragraph are positive, and the corresponding eigenvalue (which we here denote by Λ) is also positive (why?). We shall prove the following proposition: *if a vector* $\mathbf{x}_1$ *having at least one positive element satisfies the relation* $\mathbf{A}\mathbf{x}_1 \geqslant \alpha \mathbf{x}_1$ *then* $\Lambda \geqslant \alpha$. Indeed, let k be the greatest value for which $\mathbf{x}_0 \geqslant k\mathbf{x}_1$. Then $\mathbf{A}\mathbf{x}_0 \geqslant \mathbf{A}(k\mathbf{x}_1)$, i.e. $\Lambda \mathbf{x}_0 \geqslant k\alpha \mathbf{x}_1$ and $\mathbf{x}_0 \geqslant \left(\frac{\alpha}{\Lambda} k\right) \mathbf{x}_1$. Consequently, inequality $\alpha > \Lambda$ contradicts the definition of k whence $\Lambda \geqslant \alpha$. We can similarly prove that *if all the elements of a vector* $\mathbf{x}_2$ *are positive and it satisfies the relation* $\beta \mathbf{x}_2 \geqslant \mathbf{A}\mathbf{x}_2$ *then* $\beta \geqslant \Lambda$.

The former proposition immediately shows that the real eigenvalues of the matrix $\mathbf{A}$ cannot exceed Λ and the latter implies that *every eigenvector* $\mathbf{x} \geqslant \mathbf{0}$ *corresponds to the eigenvalue* Λ. It turns out that under the above assumptions the following strengthened assertions (whose proofs we do not present here) are true: *the eigenvalue* Λ *is simple*; *the moduli of the other eigenvalues of the matrix* $\mathbf{A}$ (including the pure imaginary ones provided they exist) *are smaller than the modulus of* Λ. It follows that for $\Lambda < 1$ the equation $\mathbf{x} = \mathbf{A}\mathbf{x} + \boldsymbol{\delta}$ has exactly one solution, and if $\boldsymbol{\delta} \geqslant \mathbf{0}$ then we also have $\mathbf{x} \geqslant \mathbf{0}$ (this is proved by the iteration method; e.g. see [107], Sec. XVII.18).

The inequalities $\Lambda \geqslant \alpha$ and $\Lambda \leqslant \beta$ proved above make it possible to determine upper and lower bounds for Λ. For instance, putting $\mathbf{x}_1 = \mathbf{x}_2 = (1, 1, \ldots, 1)^*$ and choosing the values of α and β in an appropriate manner we arrive at the estimation

$$\min_{1 \leqslant j \leqslant n} \sum_{k=1}^{n} a_{jk} \leqslant \Lambda \leqslant \max_{1 \leqslant j \leqslant n} \sum_{k=1}^{n} a_{jk}$$

(let the reader prove this result and also deduce the inequality $\Lambda \geqslant \max\limits_{1 \leqslant j \leqslant n} \alpha_{jj}$ by choosing another vector $\mathbf{x}_1$ in an appropriate way).

A class narrower than that of matrices with nonnegative elements is formed by the so-called *oscillatory matrices*. A matrix $\mathbf{A}$ is called **oscillatory** if all its minors are nonnegative and one of the matrices $\mathbf{A}$, $\mathbf{A}^2$, $\mathbf{A}^3$, ... has all its minors positive (by the way, the latter requirement can be replaced by the equivalent conditions that $\det \mathbf{A} > 0$, and all the elements of the form $a_{j, j+1}$ and $a_{j+1, j}$ are positive). At first glance it may seem that this property is difficult to check but it turns out that it is sometimes implied by theoretical considerations, and in this connection oscillatory matrices are encountered in some problems of the theory of vibrations. All the eigenvalues of an oscillatory matrix are positive and simple; the eigenvectors corresponding to them have an increasing number of changes of the sign of their elements, which resembles the behaviour of the sequence of the functions $\sin kx$ $(0 \leqslant x \leqslant \pi, k = 1, 2, \ldots, n)$. For the properties and applications of oscillatory matrices we refer the reader to [44] and [45].

9. Method of A. N. Krylov. Here we shall consider the method of A. N. Krylov for finding the explicit expression of the characteristic polynomial $\det (\mathbf{A} - \lambda \mathbf{I})$ of a matrix $\mathbf{A}$. If the matrix $\mathbf{A}$ is of order n then

$$\det (\mathbf{A} - \lambda \mathbf{I}) \equiv (-1)^n \lambda^n + a_1 \lambda^{n-1} + a_2 \lambda^{n-2} + \ldots + a_n$$

and it is required to determine the coefficients $a_1, a_2, \ldots, a_n$. By Hamilton-Cayley theorem (Sec. 3.3) we have

$$(-1)^n \mathbf{A}^n + a_1 \mathbf{A}^{n-1} + a_2 \mathbf{A}^{n-2} + \ldots + a_{n-1} \mathbf{A} + a_n \mathbf{I} = \mathbf{0}$$

Multiplying both sides of this relation by an arbitrary column vector $\mathbf{x}_0$ consisting of n elements and transposing the first summand on the left-hand side to the right-hand side we obtain

$$a_1 \mathbf{A}^{n-1} \mathbf{x}_0 + a_2 \mathbf{A}^{n-2} \mathbf{x}_0 + \ldots + a_{n-1} \mathbf{A} \mathbf{x}_0 + a_n \mathbf{x}_0 = (-1)^{n+1} \mathbf{A}^n \mathbf{x}_0$$

This vector equality can be written in coordinate form, which yields n equations of the first degree with n unknowns $a_1, a_2,$

..., a_n from which they are found. It should be noted that when applying this method we can avoid calculating the powers of the matrix $\mathbf{A}$ because it is sufficient to determine, in succession, the vectors $\mathbf{Ax}_0$, $\mathbf{A}^2\mathbf{x}_0 = \mathbf{A}(\mathbf{Ax}_0)$, $\mathbf{A}^3\mathbf{x}_0 = \mathbf{A}(\mathbf{A}^2\mathbf{x}_0)$, etc.

There are some modifications of the method of A. N. Krylov which are more convenient for practical computations but here we shall not dwell on them.

10. Small-Parameter Method. We shall limit ourselves to considering a variant of application of the small-parameter method. Suppose we are given a matrix $\mathbf{A}$ whose all eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and the corresponding eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are known, and it is necessary to determine, for a given matrix $\mathbf{B}$, the eigenvalues and eigenvectors of the matrix $\mathbf{A}+\varepsilon\mathbf{B}$ where ε is a small parameter. For the sake of simplicity we shall assume that all the eigenvalues are simple and consider the eigenvectors of unit length to get rid of arbitrary scalar factors.

Let us expand the kth eigenvalue $\tilde{\lambda}_k$ and the corresponding eigenvector $\tilde{\mathbf{x}}_k$ of the matrix $\mathbf{A}+\varepsilon\mathbf{B}$ (which are dependent on ε) into series in powers of ε:

$$\tilde{\lambda}_k = \lambda_k + \varepsilon\mu_k + \ldots, \quad \tilde{\mathbf{x}}_k = \mathbf{x}_k + \varepsilon\mathbf{y}_k + \ldots \tag{19}$$

For $\varepsilon = 0$ we must obtain the given quantities λ_k and $\mathbf{x}_k$, which accounts for the form of the first terms of these series. The definition of an eigenvector implies that

$$(\mathbf{A}+\varepsilon\mathbf{B})(\mathbf{x}_k+\varepsilon\mathbf{y}_k+\ldots) = (\lambda_k+\varepsilon\mu_k+\ldots)(\mathbf{x}_k+\varepsilon\mathbf{y}_k+\ldots)$$

Opening the brackets and equating coefficients in like powers of ε we obtain

$$\mathbf{Ay}_k + \mathbf{Bx}_k = \lambda_k\mathbf{y}_k + \mu_k\mathbf{x}_k, \text{ that is } (\mathbf{A}-\lambda_k\mathbf{I})\mathbf{y}_k = -\mathbf{Bx}_k + \mu_k\mathbf{x}_k \tag{20}$$

If $\mathbf{y}_k$ is considered the unknown vector quantity in the last equation (20) then, as is known from Sec. 1.4, for the solution of the equation to exist it is necessary and sufficient that the right-hand member $-\mathbf{Bx}_k + \mu_k\mathbf{x}_k$ be orthogonal to the subspace of the solutions of the adjoint homogeneous system with matrix $\mathbf{A}^* - \lambda_k^*\mathbf{I}$. Under the hypotheses assumed this is a one-dimensional space, namely that generated by an eigenvector $\mathbf{z}_k \neq \mathbf{0}$ of the matrix $\mathbf{A}^*$ corresponding to the eigenvalue λ_k^* (this eigenvector is determined uniquely to within an arbitrary nonzero factor). Therefore, the orthogonality condition takes the form

$$(-\mathbf{Bx}_k + \mu_k\mathbf{x}_k, \mathbf{z}_k) \equiv -\mathbf{z}_k^*\mathbf{Bx}_k + \mu_k\mathbf{z}_k^*\mathbf{x}_k = 0$$

whence $\mu_k = \dfrac{\mathbf{z}_k^*\mathbf{Bx}_k}{\mathbf{z}_k^*\mathbf{x}_k}$. For this value of μ_k equation (20) has the one-parametric general solution $\mathbf{y}_k = \mathbf{y}_{k0} + C\mathbf{x}_k$ where C is an arbitrary constant and $\mathbf{y}_{k0}$ is an arbitrary fixed solution of (20). The

parameter C is determined on the basis of the normalization condition $\|\tilde{\mathbf{x}}_k\| = 1$ for the vector $\tilde{\mathbf{x}}_k$ implying the relation $\mathbf{y}_k^* \mathbf{x}_k = 0$ (why?).

A simpler result is obtained when $\mathbf{A}$ is a Hermitian matrix (Sec. 1.3). For this case we take the first equality (20) and multiply it on the left by $\mathbf{x}_j^*$ $(j \neq k)$. Then, taking advantage of the orthogonality of the eigenvectors and the equality $\mathbf{x}_j^* \mathbf{A} \mathbf{y}_k = (\mathbf{y}_k^* \mathbf{A} \mathbf{x}_j)^* = (\mathbf{y}_k^* \lambda_j \mathbf{x}_j)^* = \lambda_j \mathbf{x}_j^* \mathbf{y}_k$, we conclude that $\mathbf{x}_j^* \mathbf{y}_k = \frac{\mathbf{x}_j^* \mathbf{B} \mathbf{x}_k}{\lambda_k - \lambda_j}$ whence it finally follows (check it up!) that

$$\mathbf{y}_k = \sum_{j \neq k} \frac{\mathbf{x}_j^* \mathbf{B} \mathbf{x}_k}{\lambda_k - \lambda_j} \mathbf{x}_j, \quad \mu_k = \mathbf{x}_k^* \mathbf{B} \mathbf{x}_k$$

The subsequent terms of expansions (19) are found similarly. For the case of multiple eigenvalues we refer the reader to [47].

11. Method of Continuous Extension. The method of continuous extension with respect to a parameter is directly related to the small-parameter method. Suppose that the formulation of a problem involves a parameter λ whose variation is not small and that the solution of the problem is known for a value $\lambda = \lambda_0$ of the parameter. The method consists in forming a differential equation, with λ as an independent variable, with respect to the sought-for solution. The numerical solution of the differential equation with the given initial condition for $\lambda = \lambda_0$ can be readily computed (e.g. see [107], Secs. XV. 32-34), which yields the desired result.

We shall demonstrate this method by its application to constructing the inverse of a given matrix. Let the elements of a square matrix $\mathbf{A}$ depend on λ, i.e. $\mathbf{A} = \mathbf{A}(\lambda)$. Suppose that for $\lambda = \lambda_0$ the inverse matrix $\mathbf{A}_0^{-1} = \mathbf{A}(\lambda_0)^{-1}$ has been determined (for instance, computed by Gauss' method). Differentiating the identity $\mathbf{A}(\lambda)\mathbf{A}^{-1}(\lambda) = \mathbf{I}$ with respect to λ we readily find (check it up!) that

$$\frac{d(\mathbf{A}^{-1})}{d\lambda} = -\mathbf{A}^{-1} \frac{d\mathbf{A}}{d\lambda} \mathbf{A}^{-1} \tag{21}$$

Denoting, respectively, the elements of $\mathbf{A}$ and $\mathbf{A}^{-1}$ by a_{ij} and b_{ij} and equating the expressions of the corresponding elements of the matrices on both sides of (21) we arrive at a system of n^2 differential equations with respect to the functions $b_{ij}(\lambda)$. Solving this system numerically for the given initial conditions we determine $\mathbf{A}^{-1}(\lambda)$.

If the interval of variation of λ is large we must from time to time correct the constructed solution. This can be realized as follows. Suppose that for certain λ_1 we have received, for $\mathbf{A}^{-1}(\lambda_1)$, an approximate value $\mathbf{B}_1$. Then $\Delta\mathbf{B}_1 = \mathbf{A}^{-1}(\lambda_1) - \mathbf{B}_1$ is the correc-

tion we must determine. To evaluate $\Delta \mathbf{B}_1$ we put $\mathbf{A}(\lambda_1) = \mathbf{A}_1$ (this matrix is known) and write $\mathbf{A}^{-1}(\lambda_1)\mathbf{A}_1 = \mathbf{I}$, which yields

$$(\mathbf{B}_1 + \Delta \mathbf{B}_1)\mathbf{A}_1 = \mathbf{I}, \quad \Delta \mathbf{B}_1 \cdot \mathbf{A}_1 = \mathbf{I} - \mathbf{B}_1\mathbf{A}_1$$

$$\Delta \mathbf{B}_1 = (\mathbf{I} - \mathbf{B}_1\mathbf{A}_1)\mathbf{A}_1^{-1} = (\mathbf{I} - \mathbf{B}_1\mathbf{A}_1)(\mathbf{B}_1 + \Delta \mathbf{B}_1) =$$

$$= (\mathbf{I} - \mathbf{B}_1\mathbf{A}_1)\mathbf{B}_1 + (\mathbf{I} - \mathbf{B}_1\mathbf{A}_1)\Delta \mathbf{B}_1$$

that is

$$\Delta \mathbf{B}_1 = (\mathbf{I} - \mathbf{B}_1\mathbf{A}_1)\mathbf{B}_1 + (\mathbf{I} - \mathbf{B}_1\mathbf{A}_1)\Delta \mathbf{B}_1$$

Substituting this expression of $\Delta \mathbf{B}_1$ into the right-hand side of the latter equality and performing this operation repeatedly we arrive at an iterative process which results in

$$\Delta \mathbf{B}_1 - (\mathbf{I} - \mathbf{B}_1\mathbf{A}_1)\mathbf{B}_1 + (\mathbf{I} - \mathbf{B}_1\mathbf{A}_1)^2 \mathbf{B}_1 + (\mathbf{I} - \mathbf{B}_1\mathbf{A}_1)^3 \mathbf{B}_1 + \ldots \quad (22)$$

We have $\mathbf{B}_1 \approx \mathbf{A}_1^{-1}$, i.e. $\mathbf{B}_1\mathbf{A}_1 \approx \mathbf{I}$, and therefore the series on the right-hand side of (22) converges fast.

The method of continuous extension with respect to a parameter can also be used for solving nonlinear equations. For instance, let there be given a system of equations

$$f_1(x_1, x_2, \ldots, x_n; \lambda) = 0$$

$$f_2(x_1, x_2, \ldots, x_n; \lambda) = 0, \ \ldots, \ f_n(x_1, x_2, \ldots, x_n; \lambda) = 0 \quad (23)$$

whose solution $x_1^0, x_2^0, \ldots, x_n^0$ (not necessarily unique) is known for a value $\lambda = \lambda_0$ of the parameter λ. If now λ varies, the solution also becomes variable, that is $x_i = x_i(\lambda)$ $(i = 1, 2, \ldots, n)$. Differentiating equalities (23) with respect to λ we get

$$\sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j} \frac{dx_j}{d\lambda} = -\frac{\partial f_i}{\partial \lambda} \qquad (i = 1, 2, \ldots, n)$$

whence

$$\frac{dx_i}{d\lambda} = -\sum_{j=1}^{n} b_{ij} \frac{\partial f_j}{\partial \lambda} \qquad (i = 1, 2, \ldots, n) \quad (24)$$

where $b_{ij} = b_{ij}(x_1, x_2, \ldots, x_n; \lambda)$ are the elements of the inverse matrix of $\mathbf{A} = \left(\frac{\partial f_i}{\partial x_j}\right)$. These elements satisfy the system of n^2 equations obtained from equality (21). But this system cannot be solved independently because the elements of the matrix $\mathbf{A}$ now depend on the unknown quantities x_i. But we add equations (24) to this system and thus obtain a system of $n^2 + n$ differential equations with unknown functions $x_i(\lambda)$ and $b_{ij}(\lambda)$ which can be solved numerically if for $\lambda = \lambda_0$ certain initial conditions are given.

To correct the solution in the process of continuation we can apply, to system (23), Newton's method (e.g. see [107], Sec. XII.12), for certain values of λ, taking the result of continuation as a

zero approximation. If this approximation is sufficiently close to the exact solution (which can naturally be expected), the iteration process of Newton's method converges fast even if we use the approximate values of b_{ij} after the linearization has been performed. After the values of x_i's have been corrected we can also find more precise values of b_{ij}'s by applying a formula similar to (22). Then, taking these more precise values for initial ones we can again solve the differential equations and so on.

The continuation of the solution is stopped either when for a finite value of λ one or several x_i's turn into infinity (the system of differential equations in question being nonlinear, such a situation is possible; e.g. see [107], Chapter XV, Secs. 15 and 31) or if for finite x_i we obtain $\det\left(\frac{\partial f_i}{\partial x_j}\right)=0$ because then the right-hand sides of (24) turn into infinity (why?). In the latter case the well-known sufficient condition for the existence of implicit functions (e.g. see [107], Secs. IX.13 and XII.3) is violated, and therefore the solution may split into several branches or become imaginary. Here we shall not go into particulars.

This method is very effective, and therefore it is sometimes advisable to introduce a parameter artificially even when the original system of equations is of the form

$$F_1(x_1, x_2, \ldots, x_n)=0,\ F_2(x_1, x_2, \ldots, x_n)=0,\ \ldots$$
$$\ldots,\ F_n(x_1, x_2, \ldots, x_n)=0 \qquad (25)$$

and does not involve any parameters. For this purpose it is necessary to construct a system of type (23) which is in some sense related to (25), contains a parameter, can be easily solved for one value of the parameter and becomes equivalent to (25) for some other value. For instance, let the parameter (denote it by λ again) range from 0 to 1, system (23) being solvable for $\lambda=0$ and coinciding with (25) for $\lambda=1$. Then, beginning with the known solution of the system of type (23) and extending it continuously with respect to the parameter (provided it is possible), we arrive at the sought-for solution of system (25) for $\lambda=1$. When realizing this scheme we must carefully choose the form of system (23) so that the continued solution should always remain on the branch leading to the desired solution of system (25) for $\lambda=1$ when λ passes through the intermediate values.

§ 5. LINEAR PROGRAMMING

1. Basic Problem. *The basic problem of* **linear programming** *is to find an extremum of a linear function in a domain determined by a given system of linear equalities and inequalities.* If the number of independent variables is n, this problem reduces to determi-

ning an extremum of a function

$$z = c_1x_1 + c_2x_2 + \ldots + c_nx_n \qquad (1)$$

called the **objective function** (also **cost function, utility function** or **merit criterion**) under the subsidiary conditions *(constraints)*

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n \leqslant b_i \qquad (i = 1, 2, \ldots, k) \qquad (2)$$

and

$$d_{j1}x_1 + d_{j2}x_2 + \ldots + d_{jn}x_n \geqslant e_j \qquad (j = 1, 2, \ldots, m) \qquad (3)$$

If the original problem also includes inequalities of the form $d_1x_1 + d_2x_2 + \ldots + d_nx_n \leqslant e$, they can be reduced to form (3) by multiplying both sides by -1.

Equations (2) should naturally be noncontradictory (consistent) and independent, and we must also take $k < n$ (why?). Then k variables can be expressed in terms of the other $n-k$ variables. Substituting these expressions into the right-hand side of (1) and the left-hand side of (3) we obtain an analogous problem in which all the conditions imposed on the independent variables are of the form of inequalities. Hence, the problem considered here is a particular case of the problem of a conditional extremum with *unilateral constraints (nonrestricting subsidiary conditions;* e.g. see [107], Sec. XII.11). But the application of the general scheme for solving these problems based on differentiation is inexpedient here because, as will be shown, the constraints of form (3) specify a polyhedron in the space of independent variables, and function (1) attains extrema at some of its vertices. It follows that it is sufficient, in principle, to compare the values of z at the vertices and choose the extreme values among them. But if the number of subsidiary conditions is large, the number of the vertices is large as well, and therefore this "direct" method becomes very complicated and unpractical because it involves the determination of all the vertices and the corresponding values of z. For testing vertex values of function (1) the linear programming methods yield an algorithm which leads to a considerable reduction of the amount of calculations.

The development of linear programming was connected with the application of mathematical methods to economics. Beginning with 1939, problems of that kind were systematically studied by the Soviet mathematician L.V. Kantorovich (born in 1912) and his collaborators. Beginning with 1948, these problems were independently investigated in the USA. Linear programming is now an extensively developed division of mathematics. Here we shall limit ourselves to presenting some approaches to linear programming problems. For more detail we refer

the reader to special monographs (e.g. [46], [60], [114], [115], [138], [141]).

2. **Examples.** Here we give some examples of linear programming problems. Each of them has a number of modifications but we shall dwell only on one variant.

(1) The Problem of Mixtures. Let there be some products P_1, P_2, ..., P_m each of which contains components K_1, K_2, ..., K_n. For instance, these products can be fertilizers whose components are chemical elements, or food products whose components are certain kinds of fat, proteins, carbohydrates, vitamins, etc. (in the latter case we deal with the so-called **diet problem**). Let every portion of unit weight of the product P_i ($i=1, 2, \ldots, m$) contain an amount α_{ij} of the component K_j ($j=1, 2, \ldots, n$), its cost being c_i (all the quantities involved are supposed to be measured in certain units). Suppose that we compose a mixture of the given products which must contain an amount b_j of the component K_j ($j=1, 2, \ldots, m$). Generally, this can be done in many ways. Now we pose the following problem: the mixture must be composed in such a way that the total cost of the necessary products is minimal.

To formulate the problem mathematically, let us denote by x_i ($i=1, 2, \ldots, m$) the amount of the product P_i entering into the mixture. Then function (1) is the total cost $C = \sum_{i=1}^{m} c_i x_i$ of the mixture, and the problem is to minimize this function under the conditions

$$x_i \geqslant 0 \ (i=1, \ldots, m); \quad \sum_{i=1}^{m} \alpha_{ij} x_i = b_j \qquad (j=1, \ldots, n)$$

If the amounts of some products are limited we must consider additional constraints of the form $x_i \leqslant e_i$, i.e. $-x_i \geqslant -e_i$.

(2) Transportation Problem. Let factories F_1, F_2, ..., F_m manufacture certain product, the productivity of every factory F_i ($i=1, 2, \ldots, m$) being a_i tons of the product a year. This production is delivered by railway to consumers C_1, C_2, ..., C_n, the corresponding amounts being b_1, b_2, ..., b_n tons a year. It is supposed that the balance condition $\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j$ is fulfilled, which means that the total production is equal to the total consumption. Let the distance from the factory F_i ($i=1, 2, \ldots, n$) to the consumer C_j be α_{ij} km. The problem is to distribute the output (i.e. to attach the consumers to the factories) in such a way that the total freight turn-over is minimal.

To express the conditions of the problem mathematically, we denote by x_{ij} ($i=1, 2, \ldots, m$; $j=1, 2, \ldots, n$) the amount of the

product (measured in tons per year) delivered from the factory F_i to the consumer C_j. Then function (1) (measured in ton-kilometres per year) is expressed by the formula

$$z = \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_{ij} x_{ij}$$

and the problem consists in minimizing the function under the conditions

$$x_{ij} \geqslant 0 \ (i = 1, \ldots, m; \ j = 1, \ldots, n), \ \sum_{j=1}^{n} x_{ij} = a_i \ (i = 1, \ \ldots, \ m),$$

$$\sum_{i=1}^{m} x_{ij} = b_j \ (j = 1, \ \ldots, \ n)$$

Here the unknowns are numbered by means of two indices but this is inessential because they can always be arranged in a sequence (for instance, we can put $y_1 = x_{11}$, $y_2 = x_{12}$, ..., $y_n = x_{1n}$, $y_{n+1} = x_{21}$, $y_{n+2} = x_{22}$, ..., $y_{2n} = x_{2n}$, $y_{2n+1} = x_{31}$, ..., $y_N = x_{mn}$ (where $N = mn$). Then we arrive at a special case of the problem posed in Sec. 1. (In practical computation the problem can be solved without this rearrangement.)

There can be some additional requirements which make the problem more complicated. For instance, it may be necessary to take into account the cost of transportation, the possibility of using different kinds of transportation, certain conditions limiting the traffic capacity and the like. But all these problems are of the same type.

Analogous questions arise in **activity-analysis problems** connected with management and control of industry. For instance, there can be a limited amount of resources (such as raw materials, man-power, etc.) meant for manufacturing several kinds of commodities, and it is necessary to allocate the resources so as to maximize the total profit measured according to a chosen criterion. There are many other problems of that type in mathematical economics. The proper choice of the objective function plays an essential role in these problems because it determines the sought-for *optimal solution*, that is the one giving this function an extremum (i.e. maximum or minimum) under the given conditions. For instance, in the problem of maximizing the profit in management and control of industry it may turn out that it is necessary to reorganize the industrial process, and then we must take into account this additional expenditure when choosing the utility function. It should be noted that a properly posed linear programming problem involves only one objective function.

3. Geometric Considerations. As is known from analytic geometry, an equation of the form $ax + by = c$ determines a straight

line in the x, y-plane bounding two half-planes in one of which the inequality $ax+by \geqslant c$ is fulfilled while in the other we have $ax+by \leqslant c$ (the half-planes are regarded as closed, that is we attach their boundaries to them). An equation $ax+by+cz=d$ specifies a plane in space bounding two half-spaces $ax+by+cz \geqslant d$ and $ax+by+cz \leqslant d$. Similarly, an equation $\sum_{j=1}^{n} a_j x_j = b$ determines an $(n-1)$-dimensional plane ("*hyperplane*") S in the n-dimensional Cartesian space E_n of n-tuples $x_1, \ldots, x_n$. This hyperplane cuts the space into two half-spaces it bounds, namely the half-space (P^+) specified by the inequality $f(x) \geqslant b$ and (P^-) specified by the inequality $f(x) \leqslant b$ where $x=(x_1, x_2, \ldots, x_n)$ and $f(x) = \sum_{j=1}^{n} a_j x_j$. The hyperlane (S) separates these half-spaces in the sense that if we pass along a continuous curve from any point $p \in (P^+)$ to an arbitrary point $q \in (P^-)$ the value of $f(x)$ along this path continuously varies from $f(p) \geqslant b$ to $f(q) \leqslant b$, and therefore there is at least one point on the path at which it is equal to b, and then the corresponding point x lies on (S).

If we are given two points $p=(p_1, p_2, \ldots, p_n)$ and $q=(q_1, q_2, \ldots, q_n)$ in E_n, the coordinates of the point x belonging to the (straight line) segment pq and dividing it in the ratio $\frac{px}{xq}=\lambda$ are found as in plane and solid analytic geometry:

$$x_i = \frac{p_i + \lambda q_i}{1+\lambda} \qquad (i=1, 2, \ldots, n)$$

Putting $\frac{1}{1+\lambda}=\alpha$ and $\frac{\lambda}{1+\lambda}=\beta$ we can also write

$$x_i = \alpha p_i + \beta q_i \ (i=1, 2, \ldots, n) \text{ where } 0 \leqslant \alpha, \ \beta \leqslant 1, \ \alpha+\beta=1 \quad (4)$$

A set (M) of points belonging to E_n is said to be **convex** *if every line segment connecting two arbitrary points of the set is contained in the set* (i.e. all its points belong to (M)). For instance, a plane (hyperplane) of any dimension, a half-plane or half-space, a line segment, circle (sphere), an angular region in a plane including an angle less than 180° are examples of convex sets. But an angle exceeding 180°, a set of a finite number $k \geqslant 2$ of points (and, generally, any disconnected set), the circumference of a circle (and, generally, any curved line) and every nonplanar surface are nonconvex sets. In connection with the latter examples, it should be noted that the term "convexity" is used in different senses: the lines and surfaces considered convex in the ordinary geometric meaning are not convex sets in the sense of the above definition.

The notion of a **convex set** (M) **of vectors** in a linear space is introduced similarly: such a set contains, together with any two vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ belonging to (M), every vector of the form $\alpha\boldsymbol{p}+\beta\boldsymbol{q}$ where $0\leqslant\alpha,\ \beta\leqslant 1$ and $\alpha+\beta=1$.

The intersection (common portion) *of any number of convex sets belonging to the same space is always convex.* In fact, any two points p and q entering into the common portion belong to each of the intersecting sets; these sets being convex, the line segment pq belongs to each of them, and thus to their intersection.

Now consider an arbitrary set (M) of points which is not convex in the general case. Let (M_1) be the intersection of all the convex sets containing (M) as a subset (the collection of these convex sets is not empty because it always contains at least one set, namely, the whole space). Then the set (M_1) is convex, contains (M) and is contained in every convex set for which (M) is a subset (why?). In other words, (M_1) is the smallest convex set containing (M); it is referred to as the **convex hull** of the set (M). An example of a plane convex hull is depicted in Fig. 65 which shows the convex hull (M_1) of a plane set (M) consisting of two connected components (A) and (B). If a set (M) is formed of a finite number of points $p_1, p_2, \ldots, p_k$ its convex hull (M_1) is a **convex polyhedron** whose dimension is equal to the maximum number of linearly independent vectors in the system $\overrightarrow{p_1p_2}$, $\overrightarrow{p_1p_3}, \ldots, \overrightarrow{p_1p_n}$, its vertices being some (not necessarily all!) points of (M). Conversely, every convex polyhedron is the convex hull of the set of its vertices. If (M) is the set of all points of a collection of rays with common vertex O, its convex hull (M_1) is called the **convex cone spanned by** (M) (it would be more precise to say "the semicone"). In particular, a convex hull of that kind may coincide with the entire space E_n. (What is the geometric form of (M_1) when (M) consists of a finite number of concurrent rays?)

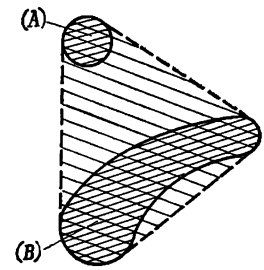


Fig. 65

Now let (M) consist of three points p, q and r; then (M_1) is a triangle which can degenerate into a line segment if the points lie in one straight line. This triangle can be obtained by joining the point r to all the points of the segment pq by means of line segments. Therefore, by (4), every point of (M_1) has coordinates of the form

$$x_i=\alpha' r_i+\beta'(\alpha p_i+\beta q_i)=\alpha_1 p_i+\alpha_2 q_i+\alpha_3 r_i \qquad (i=1,\ 2,\ \ldots,\ n)$$

where all α_j's are nonnegative and $\alpha_1+\alpha_2+\alpha_3=\beta'\alpha+\beta'\beta+\alpha'=$ $=\beta'(\alpha+\beta)+\alpha'=\beta'+\alpha'=1$.

It can be analogously shown that if (M) consists of k points $p, q, \ldots, s$, a point $x=(x_1, x_2, \ldots, x_n)$ belongs to (M_1) if and only if its coordinates are representable in the form

$$x_i=\alpha_1 p_i+\alpha_2 q_i+\ldots+\alpha_k s_i \qquad (i=1, 2, \ldots, n) \tag{5}$$

where

$$\alpha_j \geqslant 0 \ (j=1, 2, \ldots, k) \quad \text{and} \quad \alpha_1+\alpha_2+\ldots+\alpha_k=1 \tag{6}$$

But if the second condition (6) is fulfilled but α_j's are allowed to take on negative values as well the set of all points with coordinates (5) coincides with the hyperplane (S) of minimum dimension $l \leqslant k-1$ $(l \leqslant n)$ containing all the points $p, q, \ldots, s$. In particular, if $l=k-1$ (which means that among the points $p, q, \ldots, s$ there is no pair lying in a straight line, no triple lying in a plane, etc.) the parameters $\alpha_1, \alpha_2, \ldots, \alpha_k$ are uniquely determined by the point $x \in (S)$ and are termed **barycentric coordinates** of the point x; their number is equal to k but, since their sum equals unity, there are only $k-1=l$ degrees of freedom (which is in accordance with the dimension of (S)). This term is accounted for by the fact that if we place, respectively, the masses $\alpha_1, \alpha_2, \ldots, \alpha_k$ at the points $p, q, \ldots, s$, the centre of gravity of this system of material points is at x.

If $l=k-1$, the set (M_1) determined by conditions (5) and (6) is called an ***l*-dimensional simplex** (or, briefly, an ***l*-simplex**). A zero-dimensional simplex is understood as a point, a one-dimensional simplex is a closed line segment, a two-dimensional simplex is a triangle, a three-dimensional one is a tetrahedron and so on.

The convex hull of the set of all the points of a finite number of rays with common vertex O is constructed in a similar manner. If these rays are, respectively, parallel to vectors $\boldsymbol{p}, \boldsymbol{q}, \ldots, \boldsymbol{s}$ and have the same directions, the desired convex hull is obtained if we set off from O all the vectors of the form $\alpha_1\boldsymbol{p}+\alpha_2\boldsymbol{q}+\ldots$ $\ldots+\alpha_k\boldsymbol{s}$ where $\alpha_j \geqslant 0$ $(j=1, 2, \ldots, k)$.

4. Geometric Interpretation of the Basic Problem. For simplicity, let us begin with the case $n=2$ on condition that there are no constraints of form (2). Then every condition (3) determines a half-plane, and the domain (G) in which the values of function (1) are compared is the intersection of the half-planes. In Fig. 66 we see the result of intersection of five half-planes each of which is shaded in the vicinity of its bounding straight line. We see that in this example the domain (G) is the convex quadrilateral $ABCD$. There are some other cases here; for instance, if the half-plane bounded by (l_4) is replaced by the other half-plane

with the same boundary, the domain (G) becomes infinite; if this operation is performed on (l_1), the domain (G) degenerates into an **empty set** (also termed a **void** or **null set**), i.e. the one containing no points. In the latter case system of inequalities (3) is contradictory (inconsistent).

In the general case of an arbitrary dimension n the system of inequalities (3) can similarly specify a convex polyhedron, a polyhedral region receding to infinity or, finally, a void set. Generally speaking, this polyhedron or polyhedral region is n-dimensional, but there are also particular cases when the dimension is less than n; for example, if we translate the straight line (l_1) in Fig. 66 in such a way that it passes through the point B, the

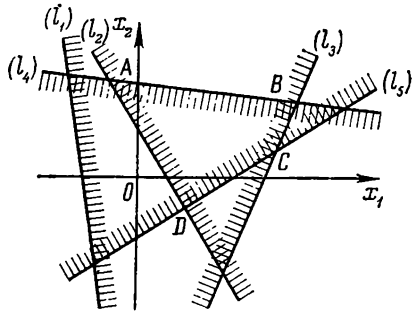


Fig 66

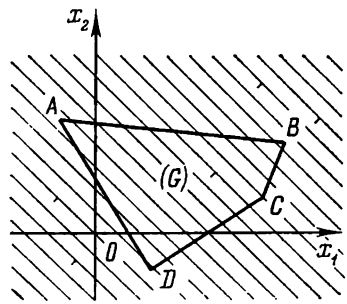


Fig. 67

corresponding system of inequalities of type (3) determines a single point, i.e. a "zero-dimensional polyhedron". Analogous conclusions are drawn in the case when there are also conditions of form (2) (provided that they are not contradictory) because they determine in E_n a hyperplane of dimension $n-\text{rank}(a_{ij})$ on which constraints (3) should be considered in a similar way.

Let us now consider the level lines of objective function (1). In Fig. 67 these level lines are depicted in the same x_1x_2-plane as in Fig. 66. The directions in which the function decreases are marked on the level lines. Function (1) being linear, the level lines form a system of parallel straight lines. If the direction of decrease of function (1) is that shown in Fig. 67, it is obvious that the greatest value of (1) is attained in the domain (G) at the vertex B, and the least value at the vertex D. By the way, if the level lines are parallel to one of the sides of the quadrilateral $ABCD$ the extremal value is attained at all the points of that side.

Similarly, for any n, the level surfaces of function (1) are mutually parallel $(n-1)$-dimensional hyperplanes. If (G) is a polyhedron then, generally speaking, function (1) assumes its greatest

and least values at two vertices (which are, in a certain sense, "extreme"). For some special directions of the level surfaces these values can be assumed at all points of some faces of (G) of dimension $q \geqslant 1$. If (G) is an unbounded polyhedral region it may happen that the objective function has no upper or lower bound (or neither) on (G); then the corresponding extremum problem has no solution.

5. Standard Form of the Basic Problem. Before solving a problem of type (1)-(3) we usually reduce it to a standard form depending on the method applied. Here we shall dwell on a method for which constraints (2) and (3) are taken in the standard form

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n = b_i \geqslant 0 \quad (i = 1, 2, \ldots, k) \tag{7}$$

and

$$x_1 \geqslant 0, \; x_2 \geqslant 0, \; \ldots, \; x_n \geqslant 0 \tag{8}$$

It is readily shown that we can always pass from constraints (2) and (3) of arbitrary type to this standard form. Inequalities (8) are usually involved in the formulation of the problem (e.g. see the examples in Sec. 2). If otherwise, we can choose any n linearly independent left members of (2) and (3), denote by $y_1, y_2, \ldots, y_n$ the differences between these left members and the corresponding right members and then pass from x's to y's. If besides inequalities of type (8) involved in the problem there is also a constraint $d_1x_1 + d_2x_2 + \ldots + d_nx_n \geqslant e$ of form (3) we put $d_1x_1 + d_2x_2 + \ldots + d_nx_n - e = x_{n+1}$ (which increases the number of independent variables by unity) and thus replace the indicated constraint by the conditions

$$d_1x_1 + d_2x_2 + \ldots + d_nx_n - x_{n+1}, = e, \quad x_{n+1} \geqslant 0$$

the former being added to system (7) and the latter to (8). If there is another constraint of form (3) we similarly introduce x_{n+2}, etc. Finally, if there is an equality of form (2) in which $b_i < 0$ we change the signs of both sides. Thus we arrive at standard constraints of forms (7) and (8).

For definiteness, we shall consider the minimization problem for function (1) under conditions (7) and (8); the maximization problem is treated analogously (by the way, we can always use the obvious relation $\max(c_1x_1 + c_2x_2 + \ldots + c_nx_n) = -\min(-c_1x_1 - -c_2x_2 - \ldots - c_nx_n)$).

The problem posed here admits of another geometric interpretation distinct from that discussed in Sec. 4. Let us introduce the k-dimensional number (column) vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ whose elements are those of the corresponding columns of the coefficient matrix (a_{ij}), the k-dimensional column vector $\mathbf{b}$ composed of the

right members b_i and the $(k+1)$-dimensional column vectors $\mathbf{a}'_j$ $(j=1, 2, \ldots, n)$ obtained from $\mathbf{a}_j$ by adding the coefficient c_j of function (1) as lowermost element. Conditions (7) are then written as the vector equality

$$x_1\mathbf{a}_1+x_2\mathbf{a}_2+\ldots+x_n\mathbf{a}_n=\mathbf{b}$$

If we make x_j's assume all the nonnegative values the vector $\mathbf{b}$ on the right-hand side describes a *cone* (cf. Sec. 3) which we referred to as *spanned by the vectors* $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$. Let it be denoted by K; this cone is a collection of vectors but we can lay off all the vectors from the origin and thus consider the cone as being a set of points belonging to E_k whose coordinates we shall denote by $\xi_1, \xi_2, \ldots, \xi_k$. System (7), (8) is noncontradictory if and only if $\mathbf{b}\in K$ (why?).

On the other hand, the set of all the vectors of the form

$$x_1\mathbf{a}'_1+x_2\mathbf{a}'_2+\ldots+x_n\mathbf{a}'_n, \quad x_j\geqslant 0 \ (j=1, 2, \ldots, n) \tag{9}$$

is a cone K', in the space E_{k+1} with coordinates $\xi_1, \xi_2, \ldots, \xi_k, z$, which is mapped on K when E_{k+1} is projected on E_k. If, in this construction, a vector belonging to set (9) is projected on $\mathbf{b}$, all relations (7), (8) are fulfilled, and hence the corresponding value $z=x_1c_1+x_2c_2+\ldots+x_nc_n$ is one of the values of (1) which are to be compared. Since we must take the least of these values, the problem is thus reduced to determining the coordinate z of the lowest (in the sense of the downward direction taken along the positive z-axis) point of intersection of the straight line $\xi_1=b_1, \xi_2=b_2, \ldots, \xi_n=b_n$ with the cone K'.

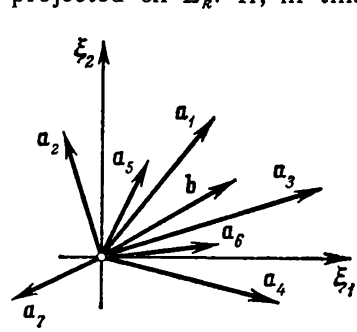


Fig. 68

6. Simplex Method. Here we shall consider a finite iterative algorithm known as the **simplex method** which is applicable to any linear programming problem taken in the above standard form. This method has an obvious geometric interpretation which we shall demonstrate by the example of the case $k=2$. Let the vectors $\mathbf{a}_j$ and $\mathbf{b}$ (see Sec. 5) in the plane E_2 be located as shown in Fig. 68. On each $\mathbf{a}_j$ the corresponding vector $\mathbf{a}'_j$ (which is not depicted in the figure) is projected. Suppose that it is possible to take two of the vectors $\mathbf{a}_j$, for instance, $\mathbf{a}_3$ and $\mathbf{a}_5$, forming a basis in E_2 in which the vector $\mathbf{b}$ has positive coordinates: $\mathbf{b}=\alpha_1\mathbf{a}_3+\beta_1\mathbf{a}_5$ where $\alpha_1>0$ and $\beta_1>0$. Then the values $x_1=x_2=x_4=x_6=x_7=0$, $x_3=\alpha_1$, $x_5=\beta_1$ satisfy all conditions (7) and (8), and consequently the corresponding value of function (1) is found

if we take the intersection of the plane (Π_1) passing through the vectors $\mathbf{a}_3'$ and $\mathbf{a}_5'$ with the straight line (l) parallel to the z-axis and drawn through the terminus of the vector $\mathbf{b}$. (We remind the reader that all the vectors are regarded as being set off from the origin.) If none of the vectors $\mathbf{a}_j'$ lies under (Π_1), the whole cone K' is above (Π_1), only some of its faces belonging to (Π_1). Hence, the value $z=\alpha_1 c_3+\beta_1 c_5$ thus found is the desired minimum value of z. But suppose now that one of the vectors $\mathbf{a}_j'$ lies under (Π_1), for instance, the vector $\mathbf{a}_1'$. Then we replace, in the basis we have chosen, the vector $\mathbf{a}_5$ by $\mathbf{a}_1$. (We have substituted $\mathbf{a}_1$ for $\mathbf{a}_5$ but not for $\mathbf{a}_3$ because in the case depicted in Fig. 68 this would lead to a negative coefficient in the resolution of the vector $\mathbf{b}$ with respect to the new basis.) The next step is to draw a plane (Π_2) passing through the vectors $\mathbf{a}_1'$ and $\mathbf{a}_3'$ and to repeat the same argument as that applied above to (Π_1). For example, if the vector $\mathbf{a}_4'$ lies under (Π_2) we pass to the basis $\mathbf{a}_1$, $\mathbf{a}_4$ in E_2 and so on. After each step the corresponding value of function (1) is decreased, and therefore, since the total number of the possible ways of choosing a basis from the given vectors $\mathbf{a}_j$ is finite, we arrive at an unimprovable result by means of a finite number of steps and thus obtain the solution of the problem.

If the problem is improperly posed we can also encounter the following case. Let, for instance, the vector $\mathbf{a}_7'$ (see Fig. 68) be found under the plane drawn through $\mathbf{a}_1'$ and $\mathbf{a}_4'$. Then we can pass neither to the basis $\mathbf{a}_1$, $\mathbf{a}_7$ nor to the basis $\mathbf{a}_4$, $\mathbf{a}_7$ without violating the condition of positivity of the coordinates of the vector $\mathbf{b}$. But in this case the straight line (l) can be continued downward into the cone K' as far as desired (why?), which indicates that the minimization problem in question has no solution.

In the case of an arbitrary k the procedure is analogous. Namely, we begin with an initial basis of E_k, chosen from the given vectors $\mathbf{a}_j$, such that all the coordinates of the vector $\mathbf{b}$ are positive; these coordinates are the values of x_j's with the corresponding numbers while the other x_j's are equal to zero. Next we replace one of the base vectors by one of the remaining $\mathbf{a}_j$'s, then repeat the operation again, etc.; this construction should be made in such a way that all the coordinates of the vector $\mathbf{b}$ are positive, and the corresponding values of function (1) decrease. Then, if the problem is properly posed we arrive, after a finite number of steps, at the sought-for solution which is the unimprovable variant of the scheme.

We cannot solely resort to geometric considerations when it is necessary to obtain the numerical solution, and therefore we now pass to the formal analytical description of this algorithm. We shall suppose that the given vector $\mathbf{b}$ cannot be resolved with respect to any system of $k-1$ vectors belonging to the set $\mathbf{a}_1$,

$\mathbf{a}_2, \ldots, \mathbf{a}_n$ (this is a **nondegeneracy condition**). It is also natural to assume that $k < n$ (why?).

Let us begin with a basis $\mathbf{a}_{i_1}, \mathbf{a}_{i_2}, \ldots, \mathbf{a}_{i_k}$, chosen from among the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$, such that all α_j's in the resolution

$$\mathbf{b} = \alpha_1 \mathbf{a}_{i_1} + \alpha_2 \mathbf{a}_{i_2} + \ldots + \alpha_k \mathbf{a}_{i_k} \tag{10}$$

are nonnegative (the question of the choice of such a basis is discussed below). Then we determine the coefficients entering into the expansions of all the vectors $\mathbf{a}_i$ with respect to this basis, that is solve the equations

$$\mathbf{a}_i = p_{i1}\mathbf{a}_{i_1} + p_{i2}\mathbf{a}_{i_2} + \ldots + p_{ik}\mathbf{a}_{i_k} \qquad (i = 1, 2, \ldots, n) \tag{11}$$

To this end, we must, for each i, solve the corresponding system of k linear equations with k unknowns. If we imagine the plane (Π_1) passing through the vectors $\mathbf{a}'_{i_1}$, $\mathbf{a}'_{i_2}, \ldots, \mathbf{a}'_{i_k}$, then the displacement of the variable point in the hyperplane E_k by the vector $\mathbf{a}_{i_j}$ determines the displacement of the corresponding point in the plane (Π_1) by c_{i_j} along the z-axis (why?). Let us evaluate the expressions

$$d_i - c_i = p_{i1}c_{i_1} + p_{i2}c_{i_2} + \ldots + p_{ik}c_{i_k} - c_i \qquad (i = 1, 2, \ldots, n) \tag{12}$$

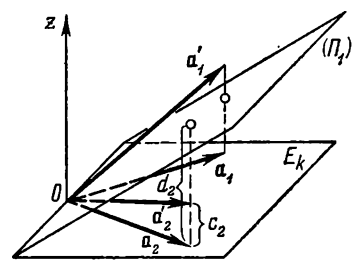


Fig. 69

whose meaning is illustrated in Fig. 69. If all these expressions are nonpositive we obtain, by the foregoing considerations, the solution of the problem which has the form $x_{i_1} = \alpha_1$, $x_{i_2} = \alpha_2, \ldots, x_{i_k} = \alpha_k$, all the other x_i's being equal to zero. This solution determines the value

$$z = z_{\min} = \alpha_1 c_{i_1} + \alpha_2 c_{i_2} + \ldots + \alpha_k c_{i_k}$$

Now suppose that one of the differences (12) is positive, say for $i = r$ (in Fig. 69 we have $r = 2$). If there are several positive differences it appears natural to take the greatest of them. For $i = r$ relations (10) and (11) then imply that for any $\theta > 0$ the equality

$$\mathbf{b} = (\alpha_1 - \theta p_{r1})\,\mathbf{a}_{i_1} + (\alpha_2 - \theta p_{r2})\,\mathbf{a}_{i_2} + \ldots + (\alpha_k - \theta p_{rk})\,\mathbf{a}_{i_k} + \theta \mathbf{a}_r \tag{13}$$

holds. Here there can be two different cases. If $p_{r1} \leqslant 0$, $p_{r2} \leqslant 0, \ldots, p_{rk} \leqslant 0$, then, for any arbitrarily large θ, all the coefficients in expansion (13) are positive, and $r \longrightarrow -\infty$ for such an expansion. This indicates that the problem under consideration has no solution.

But if there are some $p_{rj} > 0$, the values of θ cannot exceed the quantity

$$\theta = \frac{\alpha_s}{p_{rs}} = \min_{(p_{rj}>0)} \frac{\alpha_j}{p_{rj}} \tag{14}$$

(why?). For this value of θ the term with $\mathbf{a}_{i_s}$ in the right-hand side of (13) disappears whereas the other coefficients are positive. Let us replace, in the basis $\mathbf{a}_{i_1}, \mathbf{a}_{i_2}, \ldots, \mathbf{a}_{i_k}$, the vector $\mathbf{a}_{i_s}$ by $\mathbf{a}_r$ and renumber the k vectors thus obtained as $\mathbf{a}'_{i_1}, \mathbf{a}'_{i_2}, \ldots, \mathbf{a}'_{i_k}$. Then, taking equality (11) for $i = r$, we can readily express the vector $\mathbf{a}_{i_s}$ in terms of these k vectors. In particular, it follows that the vectors $\mathbf{a}'_{i_1}, \mathbf{a}'_{i_2}, \ldots, \mathbf{a}'_{i_k}$ also form a basis in E_k. Substituting the expression of $\mathbf{a}_{i_s}$ thus found into (10) and (11) we arrive at the resolutions of the form

$$\begin{gathered} \mathbf{b} = \alpha'_1 \mathbf{a}'_{i_1} + \alpha'_2 \mathbf{a}'_{i_2} + \ldots + \alpha'_k \mathbf{a}'_{i_k}, \\ \mathbf{a}_i = p'_{i1}\mathbf{a}'_{i_1} + p'_{i2}\mathbf{a}'_{i_2} + \ldots + p'_{ik}\,\mathbf{a}'_{i_k} \qquad (i = 1, 2, \ldots, n) \end{gathered} \tag{15}$$

Now we are ready to make the next step. We again form expressions of type (12) and test their signs, etc. This finite step-by-step procedure either leads to the final, i. e. unimprovable, result (usually the number of the necessary steps ranges from k to $2k$) or indicates that the problem possesses no solution.

The nondegeneracy condition introduced above does not require a special check. If it is violated this is readily recognized in the computation process because then, in expansion (15) or in some other expansion of this type, there appear zero coefficients. In this case we can make a sufficiently small variation of the given vector $\mathbf{b}$ (usually taken at random) so that the problem becomes nondegenerate and its solution can be found. The final basis thus obtained will yield an approximate solution of the original problem provided that the variation is sufficiently small.

It now remains to discuss the question of choosing the original basis $\mathbf{a}_{i_1}, \mathbf{a}_{i_2}, \ldots, \mathbf{a}_{i_k}$. In some cases it is possible to determine it in a simple manner by examining the given vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n, \mathbf{b}$ or with the aid of some considerations connected with the real significance of the problem, beginning with an intermediate solution that must be improved.

But if all this cannot be achieved, it is advisable to resort to the artificial method presented below. Let us introduce new variables $x_{n+1}, x_{n+2}, \ldots, x_{n+k}$ by writing function (1) and the subsidiary conditions in the form

$$\left.\begin{aligned} &z = c_1x_1 + c_2x_2 + \ldots + c_nx_n + \omega x_{n+1} + \omega x_{n+2} + \ldots + \omega x_{n+k} \\ &a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n + x_{n+i} = b_i \qquad (i = 1, 2, \ldots, k) \\ &x_1 \geqslant 0, \quad x_2 \geqslant 0, \ldots, x_n \geqslant 0, \quad x_{n+1} \geqslant 0, \ldots, x_{n+k} \geqslant 0 \end{aligned}\right\} \tag{16}$$

Here ω is a very large number whose exact value is not specified in the computation process and which is operated on as a formal symbol on condition that $a\omega+b>c\omega+d$ (where a, b, c and d are ordinary numbers encountered in the computations) if $a>c$ or if $a=c$ and $b>d$. Nonlinear expressions involving ω do not appear in this scheme. It is obvious that under these conditions the solution of the original problem (1), (7), (8) is equivalent to that of the new problem (16) if for the latter we obtain $x_{n+1}=x_{n+2}=\ldots=x_{n+k}=0$ in the final result (why?).

But the set of the vectors $\mathbf{a}_{n+1}, \mathbf{a}_{n+2}, \ldots, \mathbf{a}_{n+k}$ corresponding to (16) (cf. (7)) is a basis in E_k. If all b_i's are positive (this is the nondegeneracy condition) we can take these vectors as an initial basis for problem (16) (which is artificial relative to the original problem). Then we consecutively replace the base vectors following the above scheme until the final solution of the new (artificial) problem has been determined. If it turns out that all the artificial base vectors are excluded from this solution, it will simultaneously be the solution of the original problem. But if at least one artificial base vector enters into the final basis this indicates that there is no set of numbers $x_1, x_2, \ldots, x_n$ satisfying both conditions (7) and (8), that is these conditions are contradictory.

Using some of the given vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ we can sometimes construct an initial basis containing less than k artificial base vectors.

In conclusion, we mention that in the theory of linear programming the following terminology is used: a **solution** of a linear programming problem (1), (7), (8) is any set of values of x_i $(i=1, 2, \ldots, n)$ that satisfy the given linear constraints; a solution consisting of nonnegative numbers is a **feasible** solution; a solution consisting of $n-k$ x's for which the matrix of coefficients in the constraints is not singular, and otherwise consisting of zeros, is a **basic solution**; a feasible solution that minimizes linear form (1) is an **optimal solution** which, as was shown, is one of the basic feasible solutions. The simplex method is thus a standard finite iterative scheme for solving a linear programming problem by successively determining basic feasible solutions, if any exist, and testing them for optimality.

The simplex method is applicable to any properly posed linear programming problem. But for some special classes of problems (for instance, for transportation problems; see Sec. 2) there are some more convenient methods (see the books indicated in Sec. 1).

We again stress that when the number of unknowns and inequalities is not large the linear programming problem can be solved directly by testing for optimality the vertex values of the objective function. But when this number is large (in real practical

problems we usually have tens and hundreds of them or even more) this technique becomes inapplicable, and therefore it is necessary to create a new approach. It is this need that led to the development of linear programming as an independent branch of mathematics.

7. Application to Matrix Games. The theory of games is a newly developed division of mathematics dealing with the problems of finding an optimal strategy (the best behaviour) in a competitive situation involving conflict of interests when the competing individuals (or groups of individuals), understood in a broad sense and called players, pursue opposite objects (aims), and the result of the action of each "decision-maker" depends on those chosen by the others. The mathematical schemes of this theory include both games of recreation and various problems arising in many practical situations such as determining rational processes in industry and economics, combat and warfare, design of experiment and the like.

Some problems of the theory of games were for the first time studied in 1909 by E. E. Zermelo (1871-1953), a German mathematician. The fundamental results in this field were obtained in 1928 by John von Neumann. There are a number of courses devoted to the theory of games among which we mention [10], [87], [95] and [136].

Here we shall only limit ourselves to considering a special class of games, namely the so-called two-person matrix games whose scheme can be presented as follows. Let each of the players A and B independently choose a number. For instance, A chooses a number from the set of integers $1, 2, \ldots, m$ while B from $1, 2, \ldots, n$. The condition is that if A takes a number i and B a number j, the player B pays to A an amount a_{ij} (of course if $a_{ij} < 0$ then B is in fact the winner). Here (a_{ij}) is a payoff matrix which is familiar to both players. This performance can be repeated many times.

To examine characteristic features of the game let us consider the simplest example when

$$(a_{ij}) = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{17}$$

that is when each of the players independently chooses one of the two numbers 1 and 2, and if the numbers coincide A wins and if otherwise B wins, the payment being the same in all the cases. If there is only one performance (play) of the game, neither of the players has any advantage, and therefore it is impossible to speak about the best behaviour. But if the play is repeated and after every performance each player knows the choice made by the other, there appears a *competition of strategies* in

which every player tries to discover the strategy of the other. For example, if A all the time chooses the same number, or 1 and 2 in turn, or the number taken by B in the preceding play, etc., then B can discover this strategy and win all the time. But, on the other hand, A can also discover the strategy of B.

Now, suppose that B is "much stronger" than A, for instance, A is a human being and B is an electronic computer supplied with a perfectly compiled program. Then even an extremely sophisticated **pure strategy** of A, that is one in which A makes his **moves** (i. e. his particular performances) according to a determined plan uniquely specifying these moves depending on the results of the foregoing plays, is discovered by B after a sufficient number of performances. Moreover, even if A wants to "bewilder" B and tries to choose the numbers in disorder, the player B can nevertheless find a certain regularity (which may not be so distinct as before) in this strategy and win more frequently than A.

But there is a simple way to avoid losing even when the other player is very strong. For this it is sufficient that the choice in every move of the player A be made at random in such a way that the probability of either outcome is equal to 0.5. For example, the player A may toss a coin before every move and choose 1 if the coin comes up tail and 2 if it comes up head. A strategy of this kind, when every move is a random event whose probability is known in advance, is referred to as a **mixed strategy.**

Now let us return to the case of an arbitrary payoff matrix (a_{ij}) and suppose that both players decided to follow a mixed strategy, the player A taking the number i with probability p_i and B choosing the number j with probability q_j. Of course, the conditions

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{j=1}^{m} q_j = 1, \quad p_i \geqslant 0,$$

$$q_j \geqslant 0 \qquad (i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m)$$

must be fulfilled. For these strategies the average gain of the player A in one play can be computed according to the well-known rules of the probability theory (e. g. see [107], Secs. XVIII.4 and 13); it is equal to

$$v = v(p, q) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} p_i q_j$$

where $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_m)$ are the **probability vectors** determining the mixed strategies of the players.

Suppose that A tries to increase this amount by choosing an appropriate strategy while B has the opposite intention. For simplicity, assume that the strategy of each player is known to the other.

Then, for any strategy p of A, the player B follows a strategy q for which $v = \min_q v(p, q)$, and therefore the greatest average gain of the player A becomes equal to

$$v' = \max_p (\min_q v(p, q)) \tag{18}$$

Conversely, if B intends to minimize his average loss v and A opposes it, the quantity

$$v'' = \min_q (\max_p v(p, q)) \tag{19}$$

is obtained.

The fundamental **theorem of J. Neumann** states that these quantities are coincident: $v' = v'' = v$ (their common value v is referred to as the **value of the game**). It also asserts that there exists a pair of strategies (not necessarily a single one) $p = \tilde{p}$, $q = \tilde{q}$ for which

$$v' = v(\tilde{p}, \tilde{q}) = \max_p v(p, \tilde{q}) = \min_q v(\tilde{p}, q)$$

It follows (why?) that these strategies $\tilde{p}$ and $\tilde{q}$ are the best (optimal) mixed strategies for both players simultaneously: if A follows a strategy distinct from $\tilde{p}$, then, generally speaking, the player B can decrease his loss (or at least act in such a way that it is not increased), and if the strategy of B is different from $\tilde{q}$, the player A can only gain. In the case $n = m = 2$, when there are only two independent variables p_1 and q_1, a possible disposition of the level lines of the function $v(p_1, q_1)$ is shown in Fig. 70. By the way, the point $(\tilde{p}_1, \tilde{q}_1)$ may fall on the boundary of the square, and then it is not a stationary point of the function. In particular, for Example (17) we obtain $\tilde{p}_1 = \tilde{p}_2 = \tilde{q}_1 = \tilde{q}_2 = \frac{1}{2}$, $v' = 0$ (check it up!).

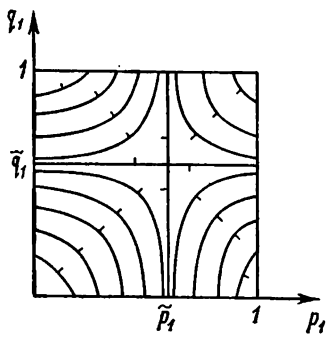


Fig. 70

One must not think that the coincidence of expressions (18) and (19) is a mere consequence of the possibility of interchanging the symbols max and min. In the general case such a permutation is illegitimate. For instance, let the reader check that

$$\max_{0 \leqslant x \leqslant 1} [\min_{0 \leqslant y \leqslant 1} (x + y - 1)^2] = 0 \quad \text{and} \quad \min_{0 \leqslant y \leqslant 1} [\max_{0 \leqslant x \leqslant 1} (x + y - 1)^2] = \frac{1}{4}$$

Consider another simple example: if i and j range from 1 to n and

$$\delta_{ij}=\begin{cases} 0 \text{ for } i\neq j, \\ 1 \text{ for } i=j \end{cases}$$

then $\max_i(\min_j \delta_{ij})=0$ and $\min_j(\max_i \delta_{ij})=1$. The proof of the Neumann theorem is in fact by far not simple and we do not present it here.

To determine the value of the game and optimal strategies, let us suppose that all $a_{ij}>0$. This can always be achieved by adding the same sufficiently large number a to all original a_{ij}'s. This addition obviously does not affect the optimal strategies and only results in an increase by a of all payoffs. Consequently, after the modified problem has been solved, we simply subtract a from the value of the game thus obtained. Furthermore, we note that the quantity

$$\min_q(\alpha_1 q_1+\alpha_2 q_2+\ldots+\alpha_m q_m)$$

is equal to the least of the numbers α_i $(i=1, 2, \ldots, m)$ for any set of numbers $\alpha_1, \alpha_2, \ldots, \alpha_m$ (let the reader prove it taking into account that $\sum_{j=1}^{m} q_j=1$). Hence, for given $p_1, p_2, \ldots, p_n$, the expression

$$\min_q v(p, q)=\min_q \sum_{j=1}^{m}\left(\sum_{i=1}^{n} a_{ij}p_i\right)q_j$$

is equal to the least of the numbers $\sum_{i=1}^{n} a_{ij}p_i$, $j=1, 2, \ldots, m$. Denoting this minimum by v_p we can clearly write $v_p>0$ and

$$\sum_{i=1}^{n} a_{ij}p_i \geqslant v_p \qquad (j=1, 2, \ldots, m) \tag{20}$$

and there is at least one value of j for which this inequality turns into equality. Putting $\frac{p_i}{v_p}=x_i$ we see that

$$x_i \geqslant 0 \qquad (i=1, 2, \ldots, n) \tag{21}$$

and

$$\sum_{i=1}^{n} a_{ij}x_i \geqslant 1 \qquad (j=1, 2, \ldots, m) \tag{22}$$

whereas

$$\sum_{i=1}^{n} x_i=\frac{1}{v_p} \tag{23}$$

Since $v_{\tilde{p}} = v'$ (why?), conditions (21), (22) are fulfilled for the values $x_i = \tilde{x}_i = \frac{\tilde{p}_i}{v'}$, and the left-hand side of (23) it equal to $\frac{1}{v'}$. On the other hand, if x_i' $(i = 1, 2, \ldots, n)$ is a solution minimizing the sum (23) under conditions (21), (22) we can compute the corresponding value v_p' by means of formula (23) and put $p_i' = x_i' v_p'$. Then we have $\sum p_i' = 1$, and, since at least one of the inequalities (22) turns into equality for $x_i = x_i'$ (why?), relations (20) and (22) imply that the quantity v_p' thus formally introduced is in fact equal to v_p and is therefore smaller than or equal to v'. It follows that $\frac{1}{v_p'} \geqslant \frac{1}{v'}$ and hence, according to the sense of the minimum problem, we conclude that $v' = v_p'$. Consequently, the value of the game is found by solving the minimum problem we have formulated here, the latter being a typical linear programming problem (see Sec. 1). The probabilities $\tilde{p}_i$ of the optimal strategy are proportional to the components $\tilde{x}_i$ of the optimal solution. The probability vector $\tilde{q}$ specifying the optimal strategy of the other player is determined in like manner.

Let us take a numerical example. Consider the payoff matrix (a_{ij}) of the form

$$\begin{pmatrix} 1 & 0 & -1 \\ -2 & 3 & 1 \end{pmatrix} \text{ (i. e. } n = 2, \ m = 3)$$

Adding 3 to all the elements we arrive at the minimum problem for the sum $x_1 + x_2$ under the conditions

$$x_1 \geqslant 0, \quad x_2 \geqslant 0, \quad 4x_1 + x_2 \geqslant 1, \quad 3x_1 + 6x_2 \geqslant 1, \quad 2x_1 + 4x_2 \geqslant 1$$

Passing to the standard form (see Sec. 5) we introduce the new variables x_3, x_4, x_5 and write

$$4x_1 + x_2 - x_3 = 1, \quad 3x_1 + 6x_2 - x_4 = 1, \quad 2x_1 + 4x_2 - x_5 = 1,$$
$$x_1 \geqslant 0, \quad x_2 \geqslant 0, \quad x_3 \geqslant 0, \quad x_4 \geqslant 0, \quad x_5 \geqslant 0$$

According to the scheme of the simplex method (Sec. 6), we put

$$\mathbf{a}_1 = \begin{pmatrix} 4 \\ 3 \\ 2 \end{pmatrix}, \quad \mathbf{a}_2 = \begin{pmatrix} 1 \\ 6 \\ 4 \end{pmatrix}, \quad \mathbf{a}_3 = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{a}_4 = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}, \quad \mathbf{a}_5 = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix},$$

$$\mathbf{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

As is indicated in Sec. 6, we now introduce the artificial basis

$$\mathbf{a}_6=\begin{pmatrix}1\\0\\0\end{pmatrix},\quad \mathbf{a}_7=\begin{pmatrix}0\\1\\0\end{pmatrix},\quad \mathbf{a}_8=\begin{pmatrix}0\\0\\1\end{pmatrix}$$

and replace the original objective function by $z=x_1+x_2+\omega x_6+$ $+\omega x_7+\omega x_8$. Let us put down the corresponding expansions of forms (10) and (11) and also expressions of type (12):

$$\mathbf{b}=\mathbf{a}_6+\mathbf{a}_7+\mathbf{a}_8$$

$$\mathbf{a}_1=4\mathbf{a}_6+3\mathbf{a}_7+2\mathbf{a}_8,\quad d_1-c_1=4\omega+3\omega+2\omega-1=9\omega-1>0$$

$$\mathbf{a}_2=\mathbf{a}_6+6\mathbf{a}_7+4\mathbf{a}_8,\quad d_2-c_2=\omega+6\omega+4\omega-1=11\omega-1>0$$

$$\mathbf{a}_3=-\mathbf{a}_6,\quad d_3-c_3=-\omega<0$$

$$\mathbf{a}_4=-\mathbf{a}_7,\quad d_4-c_4=-\omega<0$$

$$\mathbf{a}_5=-\mathbf{a}_8,\quad d_5-c_5=-\omega<0$$

The greatest value of d_i-c_i is obtained for $i=2$, and therefore we introduce $\mathbf{a}_2$ as a new base vector. The least value of (14) being attained for $i=7$ (why?), we exclude from the basis the vector $\mathbf{a}_7$. But we have $\mathbf{a}_7=\frac{1}{6}\mathbf{a}_2-\frac{1}{6}\mathbf{a}_6-\frac{2}{3}\mathbf{a}_8$, and hence

$$\mathbf{b}=\frac{1}{6}\mathbf{a}_2+\frac{5}{6}\mathbf{a}_6+\frac{1}{3}\mathbf{a}_8$$

$$\mathbf{a}_1=\frac{1}{2}\mathbf{a}_2+\frac{7}{2}\mathbf{a}_6,\quad d_1'-c_1=-\frac{1}{2}+\frac{7}{2}\omega$$

$$\mathbf{a}_3=-\mathbf{a}_6,\quad d_3'-c_3=-\omega$$

$$\mathbf{a}_4=-\frac{1}{6}\mathbf{a}_2+\frac{1}{6}\mathbf{a}_6+\frac{2}{3}\mathbf{a}_8,\quad d_4'-c_4=-\frac{1}{6}+\frac{5}{6}\omega$$

$$\mathbf{a}_5=-\mathbf{a}_8,\quad d_5'-c_5=-\omega$$

$$\mathbf{a}_7=\frac{1}{6}\mathbf{a}_2-\frac{1}{6}\mathbf{a}_6-\frac{2}{3}\mathbf{a}_8,\quad d_7'-c_7=\frac{1}{6}-\frac{11}{6}\omega$$

The next step is (check it up!) to introduce the vector $\mathbf{a}_1$ into the basis and exclude $\mathbf{a}_6$ from it. Since $\mathbf{a}_6=\frac{2}{7}\mathbf{a}_1-\frac{1}{7}\mathbf{a}_2$, we obtain

$$\mathbf{b}=\frac{5}{21}\mathbf{a}_1+\frac{1}{21}\mathbf{a}_2+\frac{1}{3}\mathbf{a}_8$$

$$\mathbf{a}_3=-\frac{2}{7}\mathbf{a}_1+\frac{1}{7}\mathbf{a}_2,\quad d_3''-c_3=-\frac{1}{7}$$

$$\mathbf{a}_4=\frac{1}{21}\mathbf{a}_1-\frac{4}{21}\mathbf{a}_2+\frac{2}{3}\mathbf{a}_8,\quad d_4''-c_4=-\frac{1}{7}+\frac{2}{3}\omega$$

$$\mathbf{a}_5=-\mathbf{a}_8,\quad d_5''-c_5=-\omega$$

$$\mathbf{a}_6=\frac{2}{7}\mathbf{a}_1-\frac{1}{7}\mathbf{a}_2,\quad d_6''-c_6=\frac{1}{7}-\omega$$

$$\mathbf{a}_7=-\frac{1}{21}\mathbf{a}_1+\frac{4}{21}\mathbf{a}_2-\frac{2}{3}\mathbf{a}_8,\quad d_7''-c_7=\frac{1}{7}-\frac{5}{3}\omega$$

Now we replace $\mathbf{a}_8$ by $\mathbf{a}_4$, and the final basis thus found does not involve the artificial vectors which therefore are no longer of interest. We have $\mathbf{a}_8 = -\frac{1}{14}\mathbf{a}_1 + \frac{2}{7}\mathbf{a}_2 + \frac{3}{2}\mathbf{a}_4$, and consequently

$$\mathbf{b} = \frac{3}{14}\mathbf{a}_1 + \frac{1}{7}\mathbf{a}_2 + \frac{1}{2}\mathbf{a}_4$$

$$\mathbf{a}_3 = -\frac{2}{7}\mathbf{a}_1 + \frac{1}{7}\mathbf{a}_2, \quad d_3''' - c_3 = -\frac{1}{7}$$

$$\mathbf{a}_5 = \frac{1}{14}\mathbf{a}_1 - \frac{2}{7}\mathbf{a}_2 - \frac{3}{2}\mathbf{a}_4, \quad d_5''' - c_5 = -\frac{1}{7}$$

The differences of type (12) thus obtained are negative, which means that we have received the optimal solution $x_1 = \frac{3}{14}$, $x_2 = \frac{1}{7}$, $x_3 = 0$, $x_4 = \frac{1}{2}$, $x_5 = 0$. Applying formula (23) and subtracting the number 3 we determine the value of the game and the probabilities specifying the optimal strategy:

$$v' = \frac{1}{x_1 + x_2} - 3 = -\frac{1}{5} \quad \text{and} \quad \tilde{p}_1 = \frac{3}{5}, \quad \tilde{p}_2 = \frac{2}{5}$$

Let the reader show that for the other player the probabilities of his optimal strategy are $\tilde{q}_1 = \frac{2}{5}$, $\tilde{q}_2 = 0$, $\tilde{q}_3 = \frac{3}{5}$. To this end one can either independently solve the maximum problem for the sum $y_1 + y_2 + y_3$ under the conditions

$$y_1 \geqslant 0, \ y_2 \geqslant 0, \ y_3 \geqslant 0, \ 4y_1 + 3y_2 + 2y_3 \leqslant 1, \ y_1 + 6y_2 + 4y_3 \leqslant 1 \quad (24)$$

or simply put $y_1 + y_2 + y_3 = \left(3 - \frac{1}{5}\right)^{-1} = \frac{5}{14}$ and take into account that for the optimal solution at least one of inequalities (24) turns into equality. The former technique is more complicated, but in its application we incidentally check up the value of the game.

Thus, this game is unfavourable for the first player. If the second player takes an urn with two white and three black balls, draws at random a ball from it, before every move, and then chooses the number 1 if the selected ball is white and the number 3 if the selected ball is black, nobody can oppose this strategy.

8. Some Other Problems. There are various classes of problems whose formulation is close to that of the basic problem of linear programming (Sec. 1), which are not reducible to it. Although these problems involve more complicated methods than the above they can in many cases be successfully solved. Here we shall only mention the terminology used in this division of mathematics. For more detail concerning the statements of the problems and methods of solution we refer the reader to [52], [75] and [141].

There is a class of problems related to **discrete** (linear) **programming** in which all or some unknown quantities x_j can only assume

the values belonging to a given discrete set of numbers, for instance, integral values. If the expected values of the sought-for quantities are sufficiently large it is often possible to apply the following technique: we first solve the problem without taking into account that the unknown values are integers and then round off the values of the parameters in the solution thus found to the nearest integers. But if the unknown values are not large this method may prove to be inapplicable. In the books indicated above the reader can find some algorithms for solving these problems but it should be noted that their efficiency is still insufficient.

When some parameters entering into the formulation of a linear programming problem (i. e. coefficients in the objective function and in the constraints) are variable, and it is necessary to investigate the corresponding variation of the optimal solution of the problem, we speak of **parametric programming**. For instance, suppose that only the coefficients of the objective function change. Then the geometric significance of the basic problem (in particular, see Fig. 67) indicates that during some period the solution remains invariable and then, at a certain moment, it turns in jump-like manner into another solution corresponding to one of the adjacent vertices of the polyhedron in which the objective function is considered; then it remains invariable for some time, after which it again changes in jump-like manner and so on.

In some cases (in particular, when we consider problems of controlling processes developing in time) it is necessary to work out a sequence of decisions, a decision taken at a certain stage of control affecting the formulation of the problem for the subsequent stages. Then we speak of **multistage decision processes** related to the so-called **dynamic programming**.

If the total number of unknowns and subsidiary conditions involved in a problem is very large, general methods of constructing the optimal solution may prove to be ineffective because the process of solution becomes too complicated and the amount of calculations exceeds the capacity of modern computers. Then it is advisable to try to divide the sought-for quantities into groups and obtain an approximate solution of the whole problem by means of solving analogous problems for these groups and inserting corrections in order to take into account interactions between the groups. This approach is related to the so-called **partition programming**. We shall demonstrate this idea by a simple example. Let it be necessary to solve the system of equations

$$\left.\begin{aligned} a_1x+b_1y+\alpha_1z+\beta_1t&=c_1\\ a_2x+b_2y+\alpha_2z+\beta_2t&=c_2\\ \alpha_3x+\beta_3y+a_3z+b_3t&=c_3\\ \alpha_4x+\beta_4y+a_4z+b_4t&=c_4 \end{aligned}\right\} \qquad (25)$$

in which the coefficients α_i and β_i are comparatively small. Then we can construct an initial approximation to the solution by putting $\alpha_i = \beta_i = 0$ and solving the two resultant systems of equations each of which involves only two unknowns. It is of course much simpler to solve these two systems than the original one. To insert the necessary corrections, we can substitute the values of the unknowns found from the simplified systems into the terms of (25) containing α_i and β_i, again solve the two new systems of equations with two unknowns and, if necessary, iterate this process repeatedly. To realize the iteration scheme, it is expedient to compute in advance the inverses of the matrices $\begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}$ and $\begin{pmatrix} a_3 & b_2 \\ a_4 & b_4 \end{pmatrix}$, which facilitates the process of solving the systems of two equations connected with each iteration.

If a problem is solved when information is incomplete, some parameters entering into the formulation of the problem can be random quantities. Problems of this type belong to **stochastic programming.**

A wide class of important problems is related to the so-called **nonlinear programming.** In these problems the objective functions or the functions specifying the constraints or both are nonlinear. In the general case these problems are very complicated. But nevertheless there are some algorithms worked out for solving some special classes of problems involving functions subject to certain restrictions although these algorithms are generally more complicated than in linear programming.

There are nonlinear programming problems for which more effective algorithms are known, namely the problems of **convex programming.** In these problems objective functions and domains in which their values are compared are supposed to be *convex*. The definition of a convex domain was given in Sec. 5.3, and the definition of a convex function reads as follows. We say that a function $f(x)$, where $x = (x_1, x_2, \ldots, x_n)$, is **convex** if for any $\alpha \geqslant 0$ and $\beta \geqslant 0$ satisfying the condition $\alpha + \beta = 1$ and any two points $x = a$ and $x = b$ the inequality

$$f(\alpha a + \beta b) \leqslant \alpha f(a) + \beta f(b) \tag{26}$$

is fulfilled. It can be easily shown that this definition is equivalent to the following property: the intersection of the surface $z = f(x)$ in the $(n+1)$-dimensional space of the variables $z, x_1, x_2, \ldots, x_n$ and any two-dimensional hyperplane parallel to the z-axis is a curve which is convex downward (on condition that the upward direction goes along the positive z-axis; e. g. cf. [107], Sec. IV.24).

It follows from (26) that if $f(a) \leqslant c$ and $f(b) \leqslant c$ then we also have $f(\alpha a + \beta b) \leqslant c$. Consequently, the set of points for which $f(x) \leqslant c$ is always convex for any c. Therefore, the basic problem

of convex programming is formulated as follows: it is necessary to find the minimum of a function $f(x)$ under the conditions $\varphi_i(x) \leqslant c_i (i = 1, 2, \ldots, k)$ where the function f and all the functions φ_i are convex. It should be noted that every linear function is convex (why?), and hence linear programming is a particular case of convex programming. The set of functions φ_i usually includes the functions $-x_i$, the corresponding values of c_i being equal to zero. Then the constraints involve the inequalities $x_i \geqslant 0$ $(i = 1, 2, \ldots, n)$.

We can readily prove by using inequality (26) that if at a point $x = a$ belonging to a convex domain (G) the function f has a local (relative) minimum (which may be attained in the interior of (G) or on its boundary), the inequality $f(x) \geqslant f(a)$ holds for all points $x \in (G)$. Consequently, the minimum is necessarily an absolute one (why?). This property essentially simplifies the construction of the optimal solution.

The quadratic function

$$f(x) = \sum_{i, j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n} b_i x_i + c$$

is a particular case of a convex function if the matrix (a_{ij}) is symmetric and all its eigenvalues are nonnegative. Problems with such quadratic objective functions involving only linear constraints are dealt with in **quadratic programming**. This class of nonlinear programming problems is most thoroughly studied.

CHAPTER V

Tensors

The notion of a *tensor* is one of the basic mathematical concepts and is widely used in mechanics, electrodynamics, relativity theory, etc. It originally appeared in the 19th century in connection with some problems of the theory of elasticity, and the related mathematical theory was systematically developed in 1886-1901 by C.G. Ricci (1853-1925), an Italian geometer, and T. Levi-Givita (1873-1942), an Italian mathematician and mechanician. The role of this new mathematical theory essentially increased after the great scientist A. Einstein (1879-1955) created in 1915-16 the general theory of relativity whose mathematical methods are completely based on the tensor calculus.

Up to now we considered physical quantities which were scalars or vectors. But there are also physical quantities of a more complicated nature. For instance, the distribution of stresses in an elastic body subject to a deformation is characterized by the density **p** of force with which one part of the body acts upon the other across an arbitrary plane (Q) (Fig. 71). But the quantity **p** varies with the direction of the plane, and hence such a state of the body is not characterized by a vector but by a more complicated quantity which is a *tensor of rank two* (whose definition is given in Sec. 2). It turns out that many other important physical quantities characterizing the state of a continuous medium are also tensors.

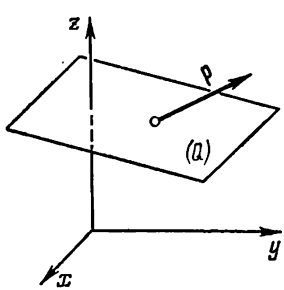


Fig. 71

Modern *tensor algebra* (studying algebraic and some other operations on tensors) and *tensor analysis* (which is the theory of tensor fields connected with the operations of differentiation and integration) are extensively developed branches of mathematics. In the present chapter we shall only dwell on some simpler facts

of the theory of tensors. For more detail we refer the reader to special monographs [14], [64], [94], [113] and [116] devoted to tensor calculus and also to the corresponding chapters of the books [50], [112] and [117] in which the fundamentals of tensor calculus are considered.

§ 1. TENSOR ALGEBRA

1. Examples. We can approach the notion of a tensor by considering the method of describing vectors in the ordinary (geometric) space with the aid of triples of numbers. As is known from vector algebra, to perform an algebraic operation on several vectors it is convenient to choose a Cartesian (Euclidean) basis $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$, resolve every vector $\mathbf{a}$ with respect to this basis in the form

$$\mathbf{a} = a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k} \tag{1}$$

and then manipulate the projections a_x, a_y, a_z, which are the numerical coefficients in resolution (1), instead of the vectors themselves. Moreover, when we consider a concrete vector it is usually more convenient to define it by means of resolution (1) than to introduce it in a purely geometric manner.

Now let us discuss the nature of the base vectors $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$. In some cases when the problem in question involves a natural frame of reference used for specifying directions in space (e. g. these are many problems of statics) the base vectors of this kind can be precisely described in terms of the data and conditions given in the formulation of the problem. But there are also many situations in which the introduction of such an "absolute" reference system looks artificial or even is impossible. Then, at first glance, it seems that we arrive at a paradox: we consider the projections of a concrete vector which depend on the choice of the basis, but we do not specify this choice.

This difficulty can be overcome if we reject the idea of taking a fixed basis and assume that all the bases are equivalent and that to every basis $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ there corresponds a set of numbers a_x, a_y, a_z, in accordance with formula (1). A set of quantities of that type which only take on certain values after a basis has been chosen and which are transformed according to certain rules when the basis is changed (see below) is referred to as a *tensor quantity* or simply a *tensor*, and the quantities themselves, which constitute the tensor, when being taken in a certain order, are called the *components* of the tensor. (Here we must note a discrepancy in the usage of terms: in vector algebra, when dealing with a vector $\mathbf{a}$, we usually apply the term "component" to the vectors $a_x\mathbf{i}$, $a_y\mathbf{j}$ and $a_z\mathbf{k}$ but not to the numbers a_x, a_y and a_z which are the projections (coordinates) of the vector $\mathbf{a}$. But in this chapter this term will be applied to the numbers a_x, a_y, a_z themselves.)

In what follows we shall write $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ instead of $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ and a_1, a_2, a_3 instead of a_x, a_y, a_z. Thus we have

$$\mathbf{a} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + a_3\mathbf{e}_3 = \sum_{i=1}^{3} a_i\mathbf{e}_i$$

In tensor algebra the summation sign $\sum$ is usually omitted (when we consider sums with respect to an index) under the so-called **summation convention** of letting the repetition of an index (subscript or superscript) denote a summation with respect to that index over its range. Therefore, we write the last formula in the form

$$\mathbf{a} = a_i\mathbf{e}_i \tag{2}$$

Here the summation index i is a **dummy index** which can be denoted by any other letter or symbol as well, that is we have $\mathbf{a} = a_i\mathbf{e}_i = a_j\mathbf{e}_j = a_k\mathbf{e}_k$, etc., the limits of summation being determined by the dimension of the space in which the tensor is considered.

Now we remind the reader of the deduction of the rule for transforming the projections (components) of vectors when the basis is changed, which is known from the theory of linear mappings (e. g. see [107], Sec. XI.7). The passage from a Euclidean basis (i. e. one consisting of mutually perpendicular unit vectors forming an orthonormal system) $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ (here and henceforward we denote it, for brevity, by $\mathbf{e}_i$) to another Euclidean basis $\mathbf{e}'_i$ (consisting of vectors $\mathbf{e}'_1$, $\mathbf{e}'_2$ and $\mathbf{e}'_3$) is carried out according to the formulas

$$\mathbf{e}'_i = \sum_{j=1}^{3} \alpha_{ij}\mathbf{e}_j \text{ or, in contracted notation, } \mathbf{e}'_i = \alpha_{ij}\mathbf{e}_j \tag{3}$$

In the last formula the letter j is a dummy index whereas i is a **free index** (which in this case can take on the values 1, 2, and 3). Writing formula (2) for the new basis and substituting (3) into it we obtain

$$\mathbf{a} = a'_i\mathbf{e}'_i = a'_i\alpha_{ij}\mathbf{e}_j$$

But since the result of the substitution must be equal to the original expression $a_j\mathbf{e}_j$ specified by (2), we have

$$a_j = \alpha_{ij}a'_i, \text{ that is } a_i = \alpha_{ji}a'_j \tag{4}$$

The transformation matrix from one Euclidean basis to another being orthogonal and therefore equal to the inverse matrix of its transpose, the solution of equalities (4) with respect to the quantities a'_i is expressed by the formulas

$$a'_i = \alpha_{ij}a_j \tag{5}$$

(let the reader carry out the calculations). Formula (5) (taken for $i = 1, 2, 3$) expresses the required transformation rule.

The application of the tensor notation reduces the amount of calculations and simplifies them (in particular, it makes them independent of the dimension of the space) but requires some experience in operating on free and dummy indices. Consider an example. Let it be required to substitute expression (5) of the coordinates a'_i into the right-hand side of (4). Then we must change the notation of the free index in (5) and denote it by j, but, for this purpose, it is necessary to replace the dummy index by any other letter different from j and not coinciding with i because i serves as free index in (4). For instance, we write $a'_j = \alpha_{jk} a_k$ whence

$$a_i = \alpha_{ji} \alpha_{jk} a_k \tag{6}$$

The German mathematician L. Kronecker (1823-1891) introduced in 1866 a special symbol δ_{ik} (called the **Kronecker delta**) which designates a tensor whose components are defined as

$$\delta_{ik} = 0 \quad (i \neq k), \quad \delta_{kk} = 1$$

Applying this symbol we can write $a_i = \delta_{ik} a_k$ (why?) whence it follows, by virtue of (6), that

$$\alpha_{ji} \alpha_{jk} = \delta_{ik}$$

This formula is equivalent to the condition that the matrix (α_{ij}) is orthogonal.

As another important example we mention here the set of nine elements of the matrix (a_{ij}) of a linear mapping A of a linear space into itself. These elements assume concrete values only after a basis $\mathbf{e}_j$ has been chosen. If $\mathbf{y} = A\mathbf{x}$, the coordinates of the vectors $\mathbf{x}$ and $\mathbf{y}$ are connected by the relations $y_i = \sum_{j=1}^{n} a_{ij} x_j$, $i = 1, 2, \ldots, n$, or, simply, $y_i = a_{ij} x_j$ (in tensor notation). Performing the transformation $\mathbf{e}'_i = \alpha_{ij} \mathbf{e}_j$ of the Euclidean basis and using the second formula (4) we get

$$\alpha_{ki} y'_k = a_{ij} \alpha_{sj} x'_s \tag{7}$$

It should be noted that formula (5) is obtained from the first formula (4) by means of a simple transfer of the coefficient α_{ij} from one side of the equality to the other. Therefore we receive from (7) the relation $y'_k = \alpha_{ki} a_{ij} \alpha_{sj} x'_s$ which implies

$$a'_{ks} = \alpha_{ki} a_{ij} \alpha_{sj}, \quad \text{that is} \quad a'_{ij} = \alpha_{ik} \alpha_{js} a_{ks}$$

In the latter formula i and j are free indices while k and s are dummy indices with respect to which the double summation is performed.

2. Cartesian Tensors. General Definition. Generalizing the examples considered in Sec. 1 we arrive at the following general definition. Let, to every Euclidean basis $\mathbf{e}_i$ of a real n-dimensional Euclidean space (R), there correspond a set of n^p real quantities (numbers) $a_{ij\ldots s}$ where $i, j, \ldots, s$ are p indices each of which can independently take on the values $1, 2, \ldots, n$. Suppose that when the transformation $\mathbf{e}'_i = \alpha_{ij}\mathbf{e}_j$ to a new Euclidean basis is made these quantities are transformed according to the formula

$$a'_{ij\ldots s} = \alpha_{ii_1}\alpha_{jj_1} \ldots \alpha_{ss_1} a_{i_1 j_1 \ldots s_1} \tag{8}$$

where the right-hand side involves a p-fold summation with respect to the dummy indices $i_1, j_1, \ldots, s_1$. Then we say that we are given a **Cartesian** (Euclidean) **tensor of rank p** (of **order p** or of **pth rank** or **order**) defined in (R), with the components $a_{ij\ldots s}$. Here the term "Cartesian" or "Euclidean" (which we shall often omit) indicates that we only deal with orthonormal bases; general bases and general tensors corresponding to them will be discussed in Sec. 6.

We thus see that, according to the above definition, the projections of a vector (see Sec. 1) form a tensor of rank one and the elements of the matrix of a linear mapping constitute a tensor of rank two. A tensor of rank zero has only one component independent of the choice of the basis and thus is a *scalar*.

To specify a concrete tensor, it is sufficient to set the values of all its components relative to one basis (and this can be done quite arbitrarily) because then its values for any other basis are found from formula (8).

There is another approach to the notion of a tensor which we shall also use. Note that usually we regard a vector not as a triple of numbers but as an expression of form (2) (even when we are not interested in its geometric meaning). Proceeding from that idea, let us introduce the formal sum

$$\mathbf{T} = a_{ij\ldots s}\mathbf{e}_i\mathbf{e}_j \ldots \mathbf{e}_s \tag{9}$$

whose coefficients $a_{ij\ldots s}$ constitute a tensor in the sense of the above definition, the combinations $\mathbf{e}_i\mathbf{e}_j \ldots \mathbf{e}_s$ being understood as formal expressions but not as "real products". This means that the quantities $a_{ij\ldots s}$ assume concrete values only after a concrete basis has been chosen and are transformed according to formula (8) when the basis is changed. If we substitute the expressions $\mathbf{e}_i = \alpha_{i_1 i}\mathbf{e}'_{i_1}$, $\mathbf{e}_j = \alpha_{j_1 j}\mathbf{e}'_{j_1}$, $\ldots$, $\mathbf{e}_s = \alpha_{s_1 s}\mathbf{e}'_{s_1}$ into the right-hand side of (9), (formally) open the brackets (without changing the order of the vector "factors") and combine similar terms we obtain

$$\mathbf{T} = a_{ij\ldots s}\alpha_{i_1 i}\alpha_{j_1 j} \ldots \alpha_{s_1 s}\mathbf{e}'_{i_1}\mathbf{e}'_{j_1} \ldots \mathbf{e}'_{s_1}$$

Now, changing the notation of the dummy indices and applying

formula (8), we receive

$$\mathbf{T} = (a_{i_1 j_1 \ldots s_1} \alpha_{i i_1} \alpha_{j j_1} \ldots \alpha_{s s_1}) \mathbf{e}'_i \mathbf{e}'_j \ldots \mathbf{e}'_s = a'_{ij \ldots s} \mathbf{e}'_i \mathbf{e}'_j \ldots \mathbf{e}'_s = \mathbf{T}'$$

where $\mathbf{T}'$ is understood as an expression of form (9) in the new basis.

Therefore, in what follows *we shall more often interpret a tensor as a sum of form* (9) *which thus is invariant under any transformation of the basis*. In other words, in this interpretation we take, instead of formula (8), the expression (9) as the definition of a tensor; when passing from one basis to another we simply substitute into (9) the expressions of the old base vectors in terms of the new ones, which automatically results in the transformation of the coefficients $a_{ij \ldots s}$ in accordance with the original formula (8). If a tensor $\mathbf{T}$ is determined by its expressions in all the bases these expressions must be coherent in the above sense, but it is also sufficient to set an expression of form (9) in an arbitrarily chosen basis because then all the other expressions for the other bases are readily found from the original expression.

For example, let a tensor of second rank defined in a two-dimensional space have the components $\begin{pmatrix} 2 & 5 \\ -3 & 0 \end{pmatrix}$ in a fixed basis $\mathbf{e}_1$, $\mathbf{e}_2$. According to (9), it can be written in this basis as

$$2\mathbf{e}_1\mathbf{e}_1 + 5\mathbf{e}_1\mathbf{e}_2 - 3\mathbf{e}_2\mathbf{e}_1 \tag{10}$$

It should be stressed that in expression (10) we cannot combine "similar" terms and simplify the expression by using ordinary rules of algebra because, as was indicated above, the expressions $\mathbf{e}_1\mathbf{e}_2$ and $\mathbf{e}_2\mathbf{e}_1$ must be understood as formal combinations but not as being equal to the scalar product $\mathbf{e}_1 \cdot \mathbf{e}_2$ of the vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ (which is equal to zero).

3. Operations on Tensors. Let us consider tensors in a space (R) whose dimension is fixed. For the sake of simplicity, we shall demonstrate the operations on tensors whose ranks are not high although the same rules apply to tensors of arbitrary ranks.

(1) Linear Operations. These are the operations of addition of tensors having the same rank and multiplication by numbers which are performed according to the rules

$$a_{ij}\mathbf{e}_i\mathbf{e}_j + b_{ij}\mathbf{e}_i\mathbf{e}_j = (a_{ij} + b_{ij})\,\mathbf{e}_i\mathbf{e}_j$$

and

$$\lambda\,(a_{ij}\mathbf{e}_i\mathbf{e}_j) = (\lambda a_{ij})\,\mathbf{e}_i\mathbf{e}_j$$

This means that when adding together tensors we perform the addition of their corresponding components and when multiplying a tensor by a scalar we multiply each component by this scalar. We thus see that the axioms of linear operations are fulfilled here,

and therefore *the set of all tensors of a given rank p is a linear space of dimension n^p in which the operations are determined by the above formulas.*

(2) Outer Product of Tensors. This operation is carried out according to the formal rule written (by definition) in the form

$$(a_{ij}\mathbf{e}_i\mathbf{e}_j)(b_{ijk}\mathbf{e}_i\mathbf{e}_j\mathbf{e}_k) = (a_{ij}\mathbf{e}_i\mathbf{e}_j)(b_{rsk}\mathbf{e}_r\mathbf{e}_s\mathbf{e}_k) = a_{ij}b_{rsk}\mathbf{e}_i\mathbf{e}_j\mathbf{e}_r\mathbf{e}_s\mathbf{e}_k \qquad (11)$$

Thus, the **outer (tensor) multiplication** of two tensors of ranks p and q results in the tensor of rank $p+q$ (called the **outer** or **tensor product** of the given tensors) whose components are found by multiplying each component of the first tensor by each component of the other. The operation can only be performed if the indices in the first and second tensors are all different; if otherwise, we must change the notation of the indices before applying the operation (cf. Example (11)). It can easily be shown (but we shall not present the proof here) that the right-hand side of (11) is in fact a tensor (i. e. is transformed according to the corresponding rules when the basis is changed).

In particular, the outer product of two vectors **a** and **b** is a tensor of second rank:

$$\mathbf{ab} = (a_i\mathbf{e}_i)(b_i\mathbf{e}_i) = a_ib_j\mathbf{e}_i\mathbf{e}_j \qquad (12)$$

This tensor is called a **dyad**. We see that each of the three summands in (10) is a dyad.

(3) Contraction. This is the operation of putting one index of a tensor equal to another and summing with respect to that index. The resultant tensor is called the **contracted tensor.** For instance, contracting a tensor of rank three with components a_{ijk} with respect to the first and third indices we obtain the tensor of rank one with components $b_j = a_{iji} = \sum_{i=1}^{n} a_{iji}$, $j = 1, 2, \ldots$. Let us check that this is in fact a tensor. Suppose that the change of the basis (of form (3)) is made. Then the new values of the components are equal to (perform the calculations!)

$$b'_j = a'_{iji} = \alpha_{ii_1}\alpha_{jj_1}\alpha_{ik_1}a_{i_1j_1k_1} = (\alpha_{ii_1}\alpha_{ik_1})\alpha_{jj_1}a_{i_1j_1k_1} =$$
$$= \delta_{i_1k_1}\alpha_{jj_1}a_{i_1j_1k_1} = \alpha_{ij_1}a_{i_1j_1i_1} = \alpha_{jj_1}b_{j_1}$$

which is obviously what we set out to prove. Formally, this operation consists in identifying two indices of the components of the given tensor. This index is then repeated twice, which means that it becomes dummy, and therefore a summation is performed with respect to it. Consequently, this always results in reducing the order of the tensor by two (which is seen in the above example).

In particular, the contracted tensor obtained from a tensor of rank two with components a_{ij} is the scalar a_{ii}. It follows that

the sum of the diagonal elements of the matrix of a linear mapping relative to a Euclidean basis is independent of the choice of the basis. This sum is referred to as the **trace** (German **Spur**) of the matrix. By the way, this assertion is obviously implied by the fact that this sum is equal to the sum of all the roots of the corresponding characteristic equation (why?).

(4) Inner Multiplication. If we take two vectors $a_i\mathbf{e}_i$ and $b_i\mathbf{e}_i$ and contract dyad (12) obtained by their outer multiplication we arrive at the scalar product (also termed *inner product*) of the original vectors. The **inner product** of any tensors of arbitrary ranks $p \geqslant 1$ and $q \geqslant 1$ is defined in like manner, the resultant tensor being of rank $p+q-2$. (The inner multiplication is also called the **composition** of two tensors and the resultant tensor is referred to as the tensor **compounded** from the two given tensors.) In particular, the inner product of a tensor of second rank by a vector is found by the formula

$$(a_{ij}\mathbf{e}_i\mathbf{e}_j)\cdot(x_i\mathbf{e}_i) = (a_{ij}x_k\mathbf{e}_i\mathbf{e}_j\mathbf{e}_k)_{k=j} = a_{ij}x_j\mathbf{e}_i$$

Consequently, this operation results in a vector which is the image of the vector $\mathbf{x}$ under the linear mapping corresponding to the tensor $a_{ij}\mathbf{e}_i\mathbf{e}_j$.

(5) Interchanging Indices. This operation consists in interchanging any two indices in a given tensor. For example, interchanging the indices i and j in the tensor $a_{ijk}\mathbf{e}_i\mathbf{e}_j\mathbf{e}_k$ we obtain the tensor $a_{kji}\mathbf{e}_i\mathbf{e}_j\mathbf{e}_k$. There are four more tensors (find them!) which can be obtained from the given tensor $a_{ijk}\mathbf{e}_i\mathbf{e}_j\mathbf{e}_k$ by interchanging the other pairs of indices. In the general case all these tensors are different. A tensor is said to be **symmetric** with respect to all indices or a group of indices if it remains unchanged when any two of these indices are interchanged. If every such permutation of indices results in the multiplication of the original tensor by -1 it is called **skew-symmetric** (**antisymmetric**) with respect to these indices. It is readily seen that for a tensor to be symmetric (skew-symmetric) it is sufficient for it to possess this property relative to one arbitrarily chosen fixed basis.

4. Tensors of Second Rank. As has been indicated, a tensor of rank two can be interpreted as the set of elements of the matrix of the corresponding linear mapping in an arbitrarily chosen Euclidean basis. Therefore, the theory of tensors of rank two is closely related to matrix theory. (By the way, note that multiplication of matrices corresponds to inner (not to outer!) multiplication of the tensors.)

Every tensor of second rank can be represented in the form of a sum of a symmetric and skew-symmetric tensors which are uniquely determined by that tensor; this representation is realized by

means of the formula

$$a_{ij} = a_{(ij)} + a_{[ij]}, \quad a_{(ij)} = \frac{1}{2}(a_{ij} + a_{ji}), \quad a_{[ij]} = \frac{1}{2}(a_{ij} - a_{ji})$$

where the tensors $a_{(ij)}\mathbf{e}_i\mathbf{e}_j$ and $a_{[ij]}\mathbf{e}_i\mathbf{e}_j$ are, respectively, the **symmetric** and **antisymmetric** parts of the original tensor $\mathbf{T} = a_{ij}\mathbf{e}_i\mathbf{e}_j$. The proof of the uniqueness of this resolution is left to the reader. Similar representations can be analogously introduced for other tensors.

For a symmetric tensor $a_{ij}\mathbf{e}_i\mathbf{e}_j$, it is always possible to choose a basis such that all the nondiagonal components a_{ij} (i. e. those for which $i \neq j$) *are equal to zero, the diagonal elements being equal to the eigenvalues of the matrix* (a_{ij}). This proposition is a direct consequence of the theory of symmetric matrices (see Sec. IV.1.6). The construction of such a basis is referred to as **reduction of a tensor to its principal axes** because these base vectors go along the principal axes of the algebraic surface of the second order described by the equation $a_{ij}x_ix_j = 1$ (Sec. IV.2.1). (Let the reader prove that this surface is independent of the choice of the basis.)

For every skew-symmetric tensor $\mathbf{T}$ *defined in the three-dimensional space there exists a vector* $\mathbf{b}$ *such that the relation* $\mathbf{T}\cdot\mathbf{x} = \mathbf{b} \times \mathbf{x}$ *holds for any vector* $\mathbf{x}$. To establish this property, it is sufficient to prove its validity for an arbitrary fixed basis because neither its left-hand member nor right-hand member is dependent on the choice of the basis. If the basis is chosen, the matrix corresponding to the tensor $\mathbf{T}$ is uniquely determined and has a skew-symmetric form

$$\begin{pmatrix} 0 & \alpha & \beta \\ -\alpha & 0 & \gamma \\ -\beta & -\gamma & 0 \end{pmatrix}$$

It follows that

$$\mathbf{T}\cdot\mathbf{x} = (\alpha x_2 + \beta x_3)\,\mathbf{e}_1 + (-\alpha x_1 + \gamma x_3)\,\mathbf{e}_2 + (-\beta x_1 - \gamma x_2)\,\mathbf{e}_3$$

But the same result is obtained if we take the vector product of the vector $\mathbf{b} = -\gamma\mathbf{e}_1 + \beta\mathbf{e}_2 - \alpha\mathbf{e}_3$ by the vector $\mathbf{x}$ (check it up!).

5. Application to Mechanics. We shall consider points and vectors in the ordinary three-dimensional (geometric) space but use tensor notation.

1. If the origin is fixed, the coordinates x_i of any given point (which are equal to the projections of its radius vector on the coordinate axes) form a tensor of rank one. The outer multiplication of this tensor by itself results in a dyad with components x_ix_j.

Now consider a material body (Ω) (which can be nonhomogeneous in the general case) with mass density ρ. Then the set of

quantities

$$\int\limits_{(\Omega)} x_i x_j \rho \, d\Omega = \int\limits_{(\Omega)} x_i x_j \, dm \qquad (dm = \rho \, d\Omega) \tag{13}$$

is a second-rank tensor because the integral is equal to the limit of the corresponding integral sums, the latter being linear combinations of the dyads $(x_i x_j)_M \mathbf{e}_i \mathbf{e}_j$ constructed for different points $M \in (\Omega)$. Contracting this tensor, multiplying (in the sense of outer product) the resultant scalar by the Kronecker delta δ_{ij} (see Sec. 1) and subtracting the original tensor from this product we obtain the symmetric tensor

$$J_{ij} = \int\limits_{(\Omega)} (\delta_{ij} x_k x_k - x_i x_j) \, dm$$

known as the **inertia tensor** of (Ω) (relative to the given origin). Its diagonal elements

$$J_{11} = \int\limits_{(\Omega)} (x_2^2 + x_3^2) \, dm, \quad J_{22} = \int\limits_{(\Omega)} (x_1^2 + x_3^2) \, dm \text{ and } J_{33} = \int\limits_{(\Omega)} (x_1^2 + x_2^2) \, dm$$

are called, respectively, **the moments of inertia of** (Ω) **about the axes** x_1, x_2 and x_3. The offdiagonal elements of this tensor taken with the opposite sign, that is the expressions

$$\int\limits_{(\Omega)} x_1 x_2 \, dm, \quad \int\limits_{(\Omega)} x_1 x_3 \, dm \text{ and } \int\limits_{(\Omega)} x_2 x_3 \, dm$$

are referred to as the **products of inertia.** The principal axes of the inertia tensor for which the products of inertia vanish are termed the **principal axes of inertia.** The moments of inertia being positive, the quadratic form $J_{ij} x_i x_j$ is positive definite (why?). Therefore, the equation $J_{ij} x_i x_j = 1$ determines the so-called **ellipsoid of inertia** introduced by A.L. Cauchy in 1827. Putting $x_2 = x_3 = 0$ in the equation of the ellipsoid of inertia we obtain the expression $J_{11} = x_1^{-2}$ for the moment of inertia about the x_1-axis. But the x_1-axis can be chosen quite arbitrarily, and hence we conclude that the moment of inertia of the material body (Ω) about an arbitrary axis passing through the origin is equal to the reciprocal of the square of the distance between the origin and the point of intersection of that axis with the ellipsoid of inertia.

These quantities are applied in mechanics in the theory of rotary motion of a rigid body. In particular, it can be shown that the principal axes of inertia have the following physical significance: if the body (Ω) is fixed at the origin and made rotate about one of the principal axes of inertia, it proceeds to be in the rotary motion about this axis (i. e. the axis of rotation does not change

its direction in space), provided that there are no external forces. The moments of inertia serve as measures of inertia for rotary motions; the products of inertia enter into formulas expressing the reactions of the constraints at the fixed points of an arbitrary axis of rotation.

2. Let the origin of coordinates be fixed in space. Consider a vector field $\mathbf{A}$ which is a homogeneous linear function of the coordinates. In other words, if $\mathbf{A}$ is regarded as a function of the radius vector $\mathbf{r}$ it satisfies the condition $\mathbf{A}(\mathbf{r}_1+\mathbf{r}_2)=\mathbf{A}(\mathbf{r}_1)+\mathbf{A}(\mathbf{r}_2)$. Then $A_i=a_{ij}x_j$ where a_{ij} is a given tensor of rank two. According to Sec. 4, this tensor can be resolved into the sum of its symmetric and antisymmetric parts whence we obtain the formula

$$\mathbf{A}=a_{(ij)}x_j\mathbf{e}_i+\mathbf{b}\times\mathbf{r}$$

for an appropriately chosen vector $\mathbf{b}$.

In particular, if $\mathbf{A}=\mathbf{v}$ where $\mathbf{v}$ is a linear field of velocities of points of a continuous medium, the second summand in (14) describes a rotation of the whole space, occupied by the medium, as a rigid body with angular speed b about the axis parallel to the vector $\mathbf{b}$ and passing through the origin. If we choose the principal axes of the tensor $a_{(ij)}\mathbf{e}_i\mathbf{e}_j$ as coordinate axes the first summand in (14) takes the form $\lambda_i x_i\mathbf{e}_i$, and thus it expresses a superposition of three motions connected with deformations along these axes.

Let us now consider the general case of a nonlinear field $\mathbf{A}$. Taking a small neighbourhood of an arbitrary point M_0 we can write the Taylor expansion

$$\mathbf{A}=\mathbf{A}_0+\left(\frac{\partial\mathbf{A}}{\partial x_j}\right)_0(x_j-x_{j0})+\ldots=\mathbf{A}_0+\left(\frac{\partial A_i}{\partial x_j}\right)_0(x_j-x_{j0})\mathbf{e}_i+\ldots$$

where the subscript zero indicates that the corresponding values of the derivatives are taken at the point M_0, and the dots designate the terms of higher order of smallness. In this case we can apply resolution (14) to the group of the first-order terms, that is regard it as an approximate expression accurate to within infinitesimals of higher order. In particular, if $\mathbf{A}=\mathbf{v}$ where $\mathbf{v}$ is a velocity field of a continuous medium we can draw the following conclusions concerning the motion of infinitesimal volumes of the medium (e.g. cf. [107], Sec. XVI.26): the motion reduces to the deformation and rotation described by the tensor $\left(\frac{\partial v_i}{\partial x_j}\right)_0\mathbf{e}_i\mathbf{e}_j$ and the translatory motion with velocity $\mathbf{v}_0$.

If $\mathbf{A}=\mathbf{u}$ where $\mathbf{u}$ is a field of relative displacements of the points of a continuous medium, the application of Taylor's formula and the corresponding resolution of type (14) imply that, to within infinitesimals of higher order, the second summand

determines a **pure rotation** of an infinitesimal neighbourhood of the point M_0 (why?). Therefore, in the problems related to the variations of the form of this neighbourhood, which are due to the state of strain connected with a **pure deformation**, the basic role is played by the symmetric tensor with components $\frac{1}{2}\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)$ (the so-called **strain tensor**) corresponding to the first summand.

In connection with Taylor's formula we also mention the tensor of rank k of the form

$$\frac{\partial^k f}{\partial x_{i_1}\partial x_{i_2}\ldots\partial x_{i_k}}\mathbf{e}_{i_1}\mathbf{e}_{i_2}\ldots\mathbf{e}_{i_k}$$

where the values of the derivatives of the function $f(x_1, x_2, \ldots, x_k)$ are taken at a fixed point. The substitution of dx_i $(i=1, 2, \ldots, n)$ for $\mathbf{e}_i$ changes this tensor into the kth differential $d^k f$.

6. General Affine Tensors. Up to now we only used Euclidean (orthonormal) bases. But there are many problems in which it is expedient to take general bases, i.e. general affine coordinate systems in space. This leads to the theory of general affine tensors which involves some new notions and facts. In this section we shall consider tensors in a general n-dimensional linear space, which means that we shall not resort to the notions of a Euclidean basis, orthogonal mapping, etc.

The new features we have mentioned appear even in the simplest case $n=1$ (a straight line). Then we deal with a basis consisting of a single vector $\mathbf{e}$ whose choice is equivalent to indicating a positive direction on the straight line and setting a unit of length. Now suppose that we pass to a new unit measure of length, for instance, twice as great as the former; then new base vector is $\mathbf{e}'=2\mathbf{e}$. It is clear that the coordinate of any point therefore becomes twice as small: $x'=\frac{1}{2}x$. A quantity of that type which changes in an "opposite manner" relative to the base vectors is called *contravariant* (the precise definition is given below). Now let us take a linear function $y=ax$ defined on a straight line. After the above change of the basis, it is expressed as $y=2ax'$, that is $y=a'x'$ where $a'=2a$. We thus see that the coefficient of the linear form changes according to the same law as the base vector. A quantity of this kind whose variation is in coherence with that of the base vectors is called *covariant*. It is the distinction between the covariant and contravariant quantities that is characteristic of the theory of general affine tensors.

Let us proceed to give the general definition of an affine tensor. For definiteness, we shall consider a tensor of rank three. As in Sec. 2, it is determined by a set of n^3 components supplied by

three indices. These components assume concrete values only after a concrete basis in space has been chosen. But now it is necessary to distinguish between **covariant indices** and **contravariant indices,** the former being written as subscripts and the latter as superscripts (they must not of course be confused with exponents of powers which, by the way, are not encountered in this chapter). For example, let the first and the third indices be covariant and the second contravariant (any combinations are possible here). Then we denote the component as $a_{i\cdot k}^{\cdot j}$. The dots in this notation indicate the "omitted" indices. Namely, the dot in front of the contravariant index j shows that it is the second index, the dot between the covariant indices i and k makes it clear that i is the first index and k is the third. But if the order of the indices is inessential we can simply write a_{ik}^{j} and the like. If all the indices of a tensor are covariant we call it a **covariant tensor** and if all the indices are contravariant we speak about a **contravariant tensor.** A tensor having both covariant and contravariant indices is called **mixed**. Thus, the tensor $a_{i\cdot k}^{\cdot j}$ we have begun with is a mixed tensor. More precisely, it is a *mixed tensor of the third rank (of rank* $1+2$ as is usually written in order to indicate the number of covariant and contravariant indices), *covariant of order (rank)* 1 *and contravariant of order (rank)* 2 (or 1*-fold covariant and* 2*-fold contravariant*).

The summation convention (Sec. 1) is now applied only to a pair of coincident indices one of which is covariant and the other contravariant. For instance, we have

$$a_{i\cdot k}^{\cdot j}b^i = a_{1\cdot k}^{\cdot j}b^1 + a_{2\cdot k}^{\cdot j}b^2 + \ldots + a_{n\cdot k}^{\cdot j}b^n$$

The resultant sum can be denoted here by $c_{\cdot k}^{j}$ or $c_{k}^{\cdot j}$ (or simply c_k^j).

The transformation law for the components, when the basis is changed, is derived as follows. Suppose that we pass from a basis $\mathbf{e}_i$ to a new basis $\mathbf{e}'_i = \alpha_{i}^{\cdot j}\mathbf{e}_j$. Then the old base vectors are expressed in terms of the new ones by the formulas $\mathbf{e}_j = \beta_{\cdot j}^{i}\mathbf{e}'_i$ where the matrix $(\alpha_{i}^{\cdot j})$ $((\beta_{\cdot j}^{i}))$ is the inverse of the transpose of $(\beta_{\cdot j}^{i})$ $((\alpha_{i}^{\cdot j}))$, which means that

$$\alpha_{i}^{\cdot j}\beta_{\cdot j}^{k} = \alpha_{\cdot j}^{i}\beta_{\cdot k}^{j} = \delta_{\cdot k}^{i} = \begin{cases} 0 & (i \neq k) \\ 1 & (i = k) \end{cases} \tag{15}$$

where $\delta_{\cdot k}^{i}$ denotes the Kronecker delta which is now considered a 1-fold covariant and 1-fold contravariant tensor of rank two (the Kronecker delta is an example of the so-called **numerical tensors** whose components remain the same in all bases). Let the reader carefully examine this relationship between the matrices $\alpha_{i}^{\cdot j}$ and $\beta_{\cdot j}^{i}$.

If a tensor has the components $a_i^{\cdot j}{}_k$ relative to the basis $\mathbf{e}_i$ then in the basis $\mathbf{e}'_i$ its components are

$$(a_i^{\cdot j}{}_k)' = \alpha_i^{\cdot i_1}\beta^{\cdot j}_{\cdot j_1}\alpha_k^{\cdot k_1}a_{i_1}^{\cdot j_1}{}_{\cdot k_1} \tag{16}$$

In other words, the matrix $(\alpha_i^{\cdot j})$ "acts upon" the covariant indices and the matrix $(\beta^{\cdot j}_{\cdot i})$ upon the contravariant indices. Formula (16) (and analogous formulas for arbitrary tensors of any ranks) serves as the definition of the notion of a **general affine tensor.**

An affine tensor can be written in the form similar to (9). For this purpose we must take, in space, the so-called **reciprocal basis** $\mathbf{e}^i$ which changes according to the formula $(\mathbf{e}^i)' = \beta^{\cdot i}_{\cdot j}\mathbf{e}^j$ when the transformation $\mathbf{e}'_i = \alpha_i^{\cdot j}\mathbf{e}_j$ is made. A basis $\mathbf{e}^i$ reciprocal to $\mathbf{e}_i$ can be constructed if we arbitrarily set the vectors $\mathbf{e}^i$ for a fixed basis $\mathbf{e}_i$ and then make them change according to the above rule when the original basis is transformed. Using the reciprocal basis we can write the tensor with components $a_i^{\cdot j}{}_k$ in the form

$$\mathbf{T} = a_i^{\cdot j}{}_k\mathbf{e}^i\mathbf{e}_j\mathbf{e}^k \tag{17}$$

It is readily seen that replacing all the components and base vectors in (17) by the primed ones we obtain the equality $\mathbf{T}' = \mathbf{T}$. (Let the reader carry out the calculations taking into account equalities (16) and (15).)

As in Sec. 1, we can easily prove that the coordinates of a vector form a contravariant tensor of rank one, i.e. a vector must be written in the form $\mathbf{x} = x^i\mathbf{e}_i$. Indeed, if $\mathbf{e}'_i = \alpha_i^{\cdot j}\mathbf{e}_j$, then $x^i\mathbf{e}_i = (x^i)'\mathbf{e}'_i = (x^i)'\alpha_i^{\cdot j}\mathbf{e}_j = (x^j)'\alpha_j^{\cdot i}\mathbf{e}_i$ whence $x^i = \alpha_j^{\cdot i}(x^j)'$ and $(x^j)' = \beta^{\cdot j}_{\cdot i}x^i$, that is $(x^i)' = \beta^{\cdot i}_{\cdot j}x^j$, which indicates that $\mathbf{x} = x^i\mathbf{e}_i$ is a contravariant tensor. In contrast to it, the coefficients of a linear form constitute a covariant tensor of rank one, and therefore a form of that kind must be written as $y = a_ix^i$. In particular, for a given function $u(x^i)$ the values of the derivatives $\frac{\partial u}{\partial x^i}$ taken at a given point (which are the coefficients of du considered as a linear form in the variables dx^i) constitute a tensor of that type $\left(\text{it is denoted by } \frac{\partial u}{\partial \mathbf{r}}\right)$. The set of the coefficients of a quadratic form is a two-fold covariant symmetric tensor of second rank while the elements of the matrix of a given linear mapping $y^i = a_j^{\cdot i}x^j$ form a mixed tensor of rank two (of rank $1+1$, covariant of order 1 and contravariant of order 1). We also obtain a mixed tensor if we take the set of the derivatives $\frac{\partial A^i}{\partial x^j}$ for a given vector field $\mathbf{A}$; this tensor is designated by the symbol $\frac{\partial \mathbf{A}}{\partial \mathbf{r}}$. Let the reader verify these assertions.

It should be noted that if the matrix $(\alpha_i^{\ j})$ is orthogonal we have $(\beta^i_{\ j}) = (\alpha_i^{\ j})$ (why?). This accounts for the fact that in Sec. 2 we did not distinguish between the covariant and contravariant indices because we only used orthogonal transformations of bases there.

The operations on general affine tensors are analogous to those performed on Cartesian tensors. Of course, we can only add together tensors in whose components the indices with the same numbers are placed at the same positions, that is both are either subscripts or superscripts. The contraction is performed with respect to indices one of which is covariant and the other contravariant. Finally, two indices can be interchanged only if both are covariant or contravariant.

7. Affine Tensors in a Euclidean Space. The definition of a general affine tensor also applies to a Euclidean space, but in this case the reciprocal basis can easily be constructed (see Sec. 6) in an explicit form. Namely, for this purpose it is sufficient to choose, for a given basis $\mathbf{e}_i$, the basis $\mathbf{e}^j$ such that

$$\mathbf{e}_i \cdot \mathbf{e}^j = \delta^j_{\ i} \tag{18}$$

This can be done as follows: every vector $\mathbf{e}^j$ is chosen so that it is perpendicular to the $(n-1)$-dimensional plane spanned by the base vectors $\mathbf{e}_i$, $(i = 1, 2, \ldots, j-1, j+1, \ldots, n)$ and its length and direction are such that $\mathbf{e}^j \cdot \mathbf{e}_j = 1$. It is evident that these conditions specify $\mathbf{e}^j$ in a unique manner (cf. Sec. IV.1.5). Relations (18) obviously imply (let the reader perform the calculations) that the basis $\mathbf{e}^i$ thus constructed is in fact reciprocal to $\mathbf{e}_i$ in the sense of Sec. 6. Besides, the symmetry of formulas (18) shows that the basis $\mathbf{e}_i$ is, in its turn, reciprocal to $\mathbf{e}^j$. This construction makes it possible to define a Euclidean basis as one reciprocal to itself (why?). Therefore, for Euclidean bases every sum of form (17) turns into an expression of form (9).

For a Euclidean space, the so-called **covariant** and **contravariant metric tensors** are of essential importance. The components of these tensors are, respectively, determined by the formulas

$$g_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j, \quad g^{ij} = \mathbf{e}^i \cdot \mathbf{e}^j \tag{19}$$

(The covariant metric tensor g_{ij} is also referred to as the **fundamental metric tensor**.) It is clear that g_{ij} (g^{ij}) is a symmetric covariant (contravariant) tensor of order two; these tensors satisfy the relation

$$g_{il} g^{ik} = \delta^k_{\ l} \tag{20}$$

In fact, the left-hand member of (20) is a mixed tensor of second rank. If we choose as $\mathbf{e}_i$ a Euclidean basis, formula (20) is readily verified. But the quantities $\delta^k_{\ l}$ themselves form a mixed tensor

of second rank (this fact was mentioned in Sec. 6; let the reader prove it), and therefore the coincidence of the left-hand and right-hand members in one basis automatically implies that they coincide in every basis. Thus, the component g^{ij} ($i, j = 1, 2, \ldots, n$) is always $\frac{1}{g}$ times the cofactor of g_{ij} in the determinant $g = \det(g_{ij})$ which has g_{ij} in the ith row and jth column.

Relation (20) indicates that, in any fixed basis, the symmetric matrices (g_{ij}) and (g^{ij}) are mutually inverse. It can readily be shown that the quadratic forms corresponding to them are positive definite. Let the reader prove it. We also leave to the reader the deduction of the formulas

$$g_{ij}\mathbf{e}^j = \mathbf{e}_i, \quad g^{ij}\mathbf{e}_j = \mathbf{e}^i \tag{21}$$

We have introduced the covariant metric tensor by putting $g_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j$ where the expression $\mathbf{e}_i \cdot \mathbf{e}_j$ is understood in the sense of the scalar product defined in the given space. But this construction is sometimes reversed, namely, if there is a given covariant symmetric tensor g_{ij} of rank two, defined in a linear space (R), such that the quadratic form $g_{ij}x^ix^j$ is positive definite, we can put

$$(x^i\mathbf{e}_i) \cdot (y^j\mathbf{e}_j) = g_{ij}x^iy^j$$

and thus define a scalar product in the space (R) which becomes a Euclidean space for which the given tensor g_{ij} is the fundamental (covariant) metric tensor.

Metric tensors (19) are used in the operations of **raising** and **lowering indices** in arbitrary tensors. For instance, if we are given a tensor $\mathbf{T} = a^i\mathbf{e}_i$ (which is a vector) we can apply formulas (21) and thus receive the relation $\mathbf{T} = a^ig_{ij}\mathbf{e}^j = (a^jg_{ji})\mathbf{e}^i = a_i\mathbf{e}^i$ where, by definition, $a_i = g_{ij}a^j$. Similarly, we have $a^i = g^{ij}a_j$, $a^{ij}_{\cdot\cdot k} = g_{jj_1}g^{kk_1}a^{ij_1}_{\cdot\cdot k_1}$, and the like. A tensor obtained from a given tensor by raising or lowering any number of indices is said to be **associated** with the given tensor. Applying the operations of raising and lowering indices we can transform tensor relations to more convenient (equivalent) relations.

8. Pseudo-Euclidean Spaces. The requirement that the quadratic form $g_{ij}x^ix^j$ (called the **metric form**) should be positive definite is equivalent to the condition $\mathbf{x} \cdot \mathbf{x} > 0$ for any $\mathbf{x} \neq \mathbf{0}$, the latter inequality being one of the axioms of scalar product. In recent years the so-called **pseudo-Euclidean spaces** have been widely applied. A pseudo-Euclidean space is a real linear space in which, instead of an ordinary scalar product, a **pseudoscalar product** is introduced for which the above axiom is rejected while all the other axioms are retained. In other words, it is no longer required that the metric form be positive definite; it can be **indefinite** (i.e. it can be a quadratic form which is neither positive nor negative definite;

see Sec. IV.2.1). Here we discuss the case of indefinite forms because, as can easily be shown, the negative definite forms do not involve any essentially new ideas.

Taking an appropriate basis we can reduce the metric form to its canonical (diagonal) form, after which the diagonal coefficients can be made equal to ± 1 by changing the lengths of the base vectors (see the beginning of Sec. VI.2.4). In such a basis the pseudoscalar product is equal to

$$(x^i\mathbf{e}_i)\cdot(y^j\mathbf{e}_j)=\lambda_1x^1y^1+\lambda_2x^2y^2+\ldots+\lambda_nx^ny^n \qquad (22)$$

where the coefficients λ_i can assume only the three values $+1$, -1 and 0. It follows that *two n-dimensional pseudo-Euclidean spaces* (R) *and* (S) *whose metric forms are of the same rank and have the same index of inertia* (*or, equivalently, the same signature*; see Sec. IV.2.2) *are isomorphic.* This means that it is possible to establish a one-to-one correspondence between the elements of (R) and (S) which preserves the linear operations and the pseudoscalar product. To this end, we can choose bases in the two spaces in such a way that formulas (22) become similar and then associate with each other the vectors having the same sets of coefficients in their resolutions with respect to these bases (cf. [107] Secs. VII.19 and VII.21). On the other hand, applying the inertia law of quadratic forms (Sec. IV.2.2.) we can easily prove that spaces for which the above characteristics of their metric forms are different cannot be isomorphic.

A feature of the spaces with indefinite metric forms is the existence of vectors $\mathbf{x}\neq 0$ for which $\mathbf{x}\cdot\mathbf{x}=0$, that is of nonzero vectors orthogonal to themselves or, which is the same, having zero **pseudolength** (the pseudolength of a vector is naturally defined as $\sqrt{\mathbf{x}\cdot\mathbf{x}}$). If some of the eigenvalues of the matrix (g_{ij}) are negative there are vectors $\mathbf{x}$ for which $\mathbf{x}\cdot\mathbf{x}<0$, i.e. those having imaginary pseudolengths. All these facts make it necessary to revise one's intuition related to Euclidean spaces.

As a basic example of a pseudo-Euclidean space we can take the set of number vectors of the Cartesian n-space E_n for which the pseudoscalar product of two vectors having coordinates x^i and y^i is introduced by means of formula (22) where the numbers of positive, negative and zero coefficients λ_i are taken in accordance with the given structure of the metric form.

From the point of view of this terminology, Euclidean spaces are a particular case of pseudo-Euclidean spaces for which the index of inertia of their metric forms coincides with the dimensions of the spaces. In the theory of relativity an essential role is played by n-dimensional pseudo-Euclidean spaces for which the index of inertia is equal to $n-1$ and the signature to $n-2$ (i.e. there are $n-1$ positive coefficients and one negative coeffi-

cient among λ_i's). These spaces are called **Lorentz spaces** after H. Lorentz (1853-1928), a noted Dutch physicist. A linear mapping of a Lorentz space into itself preserving the squares of the pseudolengths is known as a **Lorentz transformation.** (Formula $\mathbf{x}\cdot\mathbf{y}=\frac{1}{4}[(\mathbf{x}+\mathbf{y})^2-(\mathbf{x}-\mathbf{y})^2]$ shows that a linear mapping preserving the pseudolengths also preserves the pseudoscalar products and vice versa.) The role of Lorentz transformations in a Lorentz space is similar to that of motions in a Euclidean space.

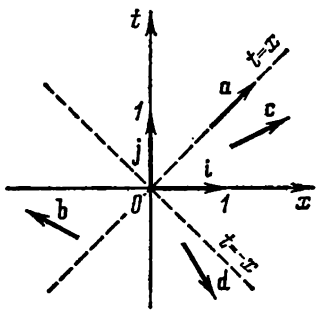


Fig. 72

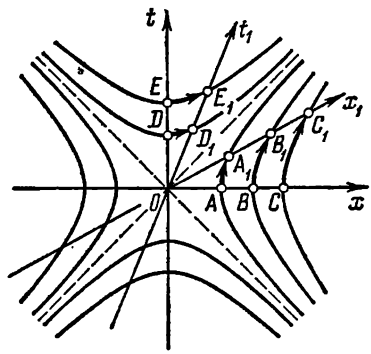


Fig. 73

Here we shall dwell in more detail on the case $n=2$ (a **Lorentz plane**). It is convenient to think of the points of such a plane as the points of an ordinary plane with coordinates x and t for which the pseudoscalar product of two vectors $\mathbf{x}_1=x_1\mathbf{i}+t_1\mathbf{j}$ and $\mathbf{x}_2=x_2\mathbf{i}+t_2\mathbf{j}$ is defined by the formula

$$(x_1\mathbf{i}+t_1\mathbf{j})\cdot(x_2\mathbf{i}+t_2\mathbf{j})=x_1x_2-t_1t_2$$

where $\mathbf{i}$ and $\mathbf{j}$ are ordinary unit vectors with origins at the point (0, 0) and termini at the points (1, 0) and (0, 1), respectively. Hence the square of the pseudolength is found by the formula

$$|x\mathbf{i}+t\mathbf{j}|^2=(x\mathbf{i}+t\mathbf{j})^2=x^2-t^2$$

In physics the quantity x is naturally interpreted as a spatial coordinate and t as time. A vector of type **a** (see Fig. 72) parallel to one of the straight lines $t=\pm x$ is of zero pseudolength. The directions of the vectors of type **b**, **c** and **i** for which the absolute value of the slope is less than unity are said to be **space-like**. The directions of the vectors of type **d** and **j** whose slopes exceed unity in their absolute values are called **time-like**. If a Lorentz transformation is written in the form $x_1=\alpha x+\beta t$, $t_1=\gamma x+\delta t$ the basic condition $x_1^2-t_1^2\equiv x^2-t^2$ together with the

additional requirements $\delta > 0$ and $\begin{vmatrix} \alpha & \beta \\ \gamma & \delta \end{vmatrix} > 0$ (the latter follow from some natural physical considerations) imply the formulas

$$x_1 = (\cosh \varphi)\, x - (\sinh \varphi)\, t, \quad x = (\cosh \varphi)\, x_1 + (\sinh \varphi)\, t_1,$$
$$t_1 = (-\sinh \varphi)\, x + (\cosh \varphi)\, t, \quad t = (\sinh \varphi)\, x_1 + (\cosh \varphi)\, t_1$$

(let the reader deduce them). Here φ is a parameter determining the transformation which is analogous to an angle of rotation. If two such transformations with parameters φ_1 and φ_2 are performed consecutively, the result is a transformation of the same type with parameter $\varphi_1 + \varphi_2$ (let the reader prove this property by means of the rule of matrix multiplication). The above transformation with parameter φ interpreted as a mapping of the xt-plane with fixed origin onto itself is called **(Lorentz) rotation through the hyperbolic angle φ**. In Fig. 73 we see the images of several points under such a transformation. The straight lines $x = \pm t$ go into themselves under the transformation. The above term is accounted for by the fact that the hyperbolas $x^2 - t^2 = C$ also go into themselves. Choosing the value of φ in an appropriate manner we can transform the x-axis in such a way that it goes in any of the space-like directions which thus are all equivalent in this theory. (In a plane of a Euclidean space we have a similar situation in the sense that all the possible directions of the x-axis are equivalent.) In these transformations the direction of the t-axis also changes. By the way, the x-axis and the t-axis are shown in Figs. 72 and 73 as mutually perpendicular but this is not important, and the axes x_1, t_1 are equivalent to them and can be used as well.

In the theory of relativity every point in the xt-plane is interpreted as an "event" occurring at time moment t at the point x (here, for simplicity, we consider "the world" as one-dimensional). The quantity t is the **proper time** in the frame of reference connected with the axis Ox, and t_1 is that in the frame of reference Ox_1, the latter being in uniform motion with velocity $\tanh \varphi$ relative to the former (why?). Since $|\tanh \varphi| < 1$, this velocity is always less than unity (in physics the quantities x and t are dimensional, and the metric form is taken as equal to the expression $x^2 - c^2 t^2$ where the constant c is the speed of light which is the maximum possible speed of a signal; then the relative speed of motion of one system with respect to another is always less than c). Formula $\tanh(\varphi_1 + \varphi_2) = \frac{\tanh \varphi_1 + \tanh \varphi_2}{1 + \tanh \varphi_1 \tanh \varphi_2}$ yields the rule for addition of relative velocities. Fig. 73 demonstrates the **relativity of simultaneity of events**; for instance, the events A, B and C are simultaneous from the point of view of an observer

connected with the frame of reference Ox but are not such for another observer moving together with the second reference system Ox_1 (let the reader find out which of these events occurs earlier than the others from the point of view of the second observer).

In a Lorentz space of dimension $n > 2$ the set of all space-like radius vectors is separated from that of the time-like vectors by the **light cone** formed of all the vectors of zero pseudolength. The former set of points is connected, that is all the space-like directions are equivalent. The latter set consists of two separate connected components; this means that *time is irreversible*, that is in any frame of reference it is impossible "to travel into the past".

9. Remark on Dimensions. In the foregoing discussion we regarded all the quantities involved as dimensionless. This is usually assumed in mathematical considerations because the theory in which we manipulate only dimensionless quantities is simpler; if the results of the theory must be applied to a concrete physical problem we most often choose some system of unit measures and thus, at the very beginning, pass to dimensionless quantities. But there are cases when such a reduction appears inconvenient, and then we should take into account the dimensions of the quantities involved.

There are two different approaches in operations on dimensional tensor quantities; this distinction is sometimes not taken into account, which leads to misunderstanding. To illustrate these approaches, consider a resolution of a dimensional vector $\mathbf{x}$ relative to a given basis $\mathbf{e}_i$: $\mathbf{x} = x^i \mathbf{e}_i$ (the case of tensors of arbitrary ranks is treated similarly). Here we can

(1) either regard the scalar (in the sense of vector algebra) quantities x^i as having the dimension of the vector $\mathbf{x}$ and the base vectors $\mathbf{e}_i$ as dimensionless

(2) or consider the quantities x_i as being dimensionless and the vectors $\mathbf{e}_i$ as having the dimension of $\mathbf{x}$.

Vector algebra and tensor algebra admit of both approaches, but one should always bear in mind the distinction between them and consistently follow the chosen interpretation.

The first approach is more universal because it enables us to consider simultaneously tensors (and, in particular, vectors) of various dimensions. The dimension of every newly introduced tensor (that is the common dimension of all its components which in this case are *physical components* of the tensor) can be quite arbitrary here. The dimension of a sum or product is found according to the general rules. But this approach involves an abstract space with dimensionless base vectors which are difficult to visualize (for instance, we then speak about a vector of length 5 but not of five units of length, etc.; on the other hand, the same

abstractness is involved when we speak about a dimensionless number 5). The metric tensor is dimensionless here, and hence the operations of raising and lowering indices in a tensor (see Sec. 6) do not affect its dimension. At the same time it follows that a sum of the form $a_i + a_i a_k a^k$ and the like are senseless.

The second approach is more visual and does not involve any other space than the space of vectors $\mathbf{x}$ but it naturally has a narrower field of application. Here we can perform the same operations as in the case of dimensionless quantities but the metric tensor becomes less convenient to manipulate. The base vectors of the reciprocal basis are of dimension $[\mathbf{x}]^{-1}$ (where $[\mathbf{x}]$ is the dimension of $\mathbf{x}$). Here we can no longer consider a tensor as a mere collection of its components and must use an expression of type (17) which thus determines the dimension of the tensor; namely, in the general case this dimension is equal to a power of the dimension of $\mathbf{x}$ whose exponent is equal to the difference between the number of contravariant indices and that of covariant indices of the components.

§ 2. TENSOR FIELDS

1. Field of a Cartesian Tensor. For definiteness, let us limit ourselves to the fields in a three-dimensional space although similar considerations are applicable to the case of an arbitrary dimension. Suppose that at each point M of the space or of its part a Cartesian tensor (Sec. 1.2)

$$\mathbf{T} = \mathbf{T}(M) = a_{ij\ldots s}(M)\,\mathbf{e}_i\,\mathbf{e}_j \ldots \mathbf{e}_s \tag{1}$$

is defined, the rank being the same for all points M. Then we say that there is a **field of a Cartesian tensor T** (further in this section we shall omit the term "Cartesian"). Scalar and vector fields are particular cases of a tensor field when the rank is, respectively, equal to zero and unity. Fields of rank two are widely applied in physics and related sciences, particularly in mechanics of continua where they have a clear physical significance (for instance, such fields are used for describing the state of strain of an elastic body; see the end of Sec. 1.5). Operations on these fields may result in the appearance of fields of higher ranks.

All the operations performed on tensors whose components are constant (see Sec. 1.3) can be applied to tensor fields. But in the case of a tensor dependent on the variable point we can also apply the operations of *tensor analysis* based on differentiation and integration. For instance, the derivative of field (1) is (by definition) computed according to the formula

$$\nabla\mathbf{T} = \frac{\partial \mathbf{T}}{\partial \mathbf{r}} = \frac{\partial a_{ij\ldots s}}{\partial x_t}\,\mathbf{e}_i\,\mathbf{e}_j \ldots \mathbf{e}_s\,\mathbf{e}_t \tag{2}$$

which generalizes, in a natural manner, the notion of the ordinary derivative of a scalar function. It can easily be shown (prove it!) that when the Euclidean basis $\mathbf{e}_i$ is changed the quantities $\frac{\partial a_{ij\ldots s}}{\partial x_t}$ (also denoted by the symbols $a_{ij\ldots s,\,t}$) are transformed in accordance with the rule of transformation of tensors (Sec. 1.2). This means that expression (2) is a tensor whose rank exceeds by unity that of (1). Derivatives (2) of a sum and product of fields are found by means of the ordinary rules of differential calculus. The differential $d\mathbf{T}$ is also expressed in terms of the derivative:

$$d\mathbf{T} = \frac{\partial \mathbf{T}}{\partial \mathbf{r}} \cdot d\mathbf{r} = a_{ij\ldots s,\,t}\, dx_t \mathbf{e}_1 \mathbf{e}_2 \ldots \mathbf{e}_s$$

The role of $d\mathbf{T}$ in tensor analysis is similar to that of ordinary differentials.

In particular, the derivative introduced above makes it possible to express the basic operations of vector analysis; namely, we have

$$\operatorname{grad} u = \frac{\partial u}{\partial \mathbf{r}} = u_{,i}\mathbf{e}_i, \quad \operatorname{div} \mathbf{A} = \mathbf{A}_{i,\,i}$$

and

$$(\mathbf{a}, \nabla)\, u = a_i u_{,i}, \quad (\mathbf{a}, \nabla)\, \mathbf{b} = a_j b_{i,\,j}\mathbf{e}_i$$

To express the rotation of a vector field it is convenient to use the notion of a **pseudotensor**. A pseudotensor differs from an ordinary tensor in its transformation rule; if the sense, the orientation, of the triple of base vectors changes, the components of a pseudotensor are additionally multiplied by -1. Now, taking the pseudotensor ε of rank 3 whose components, in any right-handed coordinate system, are determined by the relations $\varepsilon_{123} = \varepsilon_{312} = \varepsilon_{231} = 1$, $\varepsilon_{132} = \varepsilon_{213} = \varepsilon_{321} = -1$, $\varepsilon_{ijk} = 0$ when at least two of the indices i, j, k coincide, we readily show that $\operatorname{rot} \mathbf{a} = -\varepsilon_{ijk} a_{j,\,k}\mathbf{e}_i$.

By analogy with the divergence of a vector field, we introduce the notion of the divergence of a tensor field. For instance, if $a_{ij}\mathbf{e}_i\mathbf{e}_j$ is a tensor field of second rank we put

$$\operatorname{div}_{\mathrm{I}} (a_{ij}\mathbf{e}_i\mathbf{e}_j) = a_{ij,\,i}\mathbf{e}_j \quad \text{and} \quad \operatorname{div}_{\mathrm{II}} (a_{ij}\mathbf{e}_i\mathbf{e}_j) = a_{ij,\,j}\mathbf{e}_i$$

that is the rank of the resultant field is less by unity than that of the original field. We see that for nonsymmetric tensors we must additionally indicate the index with respect to which the operation of taking divergence is applied.

The integral of a tensor field taken over a line, surface or volume is introduced in a natural manner and is a tensor of the same rank (why?). The computation of a flux $\int \mathbf{T} \cdot d\boldsymbol{\sigma}$ (and, similarly, of a circulation) involves inner multiplication of the tensor by a vector

and results in a tensor (not a tensor field but a "constant" tensor) whose rank is less by unity than the original rank. If the tensor $\mathbf{T}$ is not symmetric we must additionally indicate the index with respect to which the inner multiplication (and therefore the flux) is taken.

Ostrogradsky's formula (I.1.4) is readily generalized to tensor fields. For example, let us take a tensor field $\mathbf{T}=a_{ij}\mathbf{e}_i\mathbf{e}_j$ of second rank. Considering the basis in space as fixed and applying to the vector field $a_{ij},\mathbf{e}_i$ formula (I.1.4) we obtain

$$\oint_{(\sigma)} a_{ij}n_i\,d\sigma=\int_{\Omega} a_{ij,\,i}\,d\Omega$$

Now, multiplying both sides by $\mathbf{e}_j$ and summing with respect to j, we arrive at the required formula

$$\oint_{(\sigma)} \mathbf{T}\cdot d\boldsymbol{\sigma}=\oint_{\Omega} \operatorname{div}\mathbf{T}\,d\Omega$$

in which the flux and the divergence are taken with respect to one and the same index. The final formula is irrelevant of the concrete choice of the basis and thus has a tensor sense. It holds for tensors of any rank and makes it possible to regard the divergence as the ratio of the flux of the tensor field through an infinitesimal closed surface to the volume bounded by the surface.

2. Christoffel Symbols. Now we proceed to studying tensor fields (and, in particular, vector fields) defined by means of curvilinear coordinates in a linear or Euclidean space of arbitrary dimension n. (Here and henceforward we shall use some notions related to line and surface integrals; e.g. see [107], Secs. XVI.14 and 15.) For the sake of simplicity it is advisable to interpret the results for the case $n=2$ or $n=3$ although all the considerations are the same for any dimension n.

Suppose that in a domain (G) lying in a linear space (R) some curvilinear coordinates $x^1, x^2, \ldots, x^n$ are introduced. Then, generally speaking, we can take, at each point $M\in(G)$, the vectors $\mathbf{e}_i=\left(\dfrac{\partial\mathbf{r}}{\partial x^i}\right)_M$ as base vectors. But this basis changes from point to point and thus we have a *field of bases*. If we pass to another system of curvilinear coordinates $x^{1'}, x^{2'}, \ldots, x^{n'}$, the base vectors are transformed according to the formula

$$\mathbf{e}'_i=\frac{\partial\mathbf{r}}{\partial x^{i'}}=\frac{\partial}{\partial x^j}\frac{\partial x^j}{\partial x^{i'}}=\frac{\partial x^j}{\partial x^{i'}}\mathbf{e}_j$$

(It can be easily shown that the index j in the expression $\dfrac{\partial}{\partial x^j}$ is covariant and in the expression $\dfrac{\partial}{\partial x_j}$ contravariant, and hence

the expression $\frac{\partial \mathbf{r}}{\partial x^j}\frac{\partial x^j}{\partial x^i}$ involves a summation with respect to j.) Therefore, by Sec. 1.6, the coordinates of an arbitrary vector $\mathbf{a}=a^i\mathbf{e}_i$ are transformed, under this change of the basis, according to the formula $a^{i'}=\frac{\partial x^{i'}}{\partial x^j}a^j$.

Now let us consider a problem which is important for our further aims. Suppose that a vector $\mathbf{a}$ is in a translatory motion, that is it moves in space without changing its length and direction. We can resolve this vector with respect to the above base vectors at every point it passes in the process of motion; but the basis varies from point to point and, consequently, it is clear that the projections of the vector will also be variable. Here we are going to establish the law of this variation.

The vector $\mathbf{a}=a^i\mathbf{e}_i$ being constant, we have

$$d\mathbf{a}=(da^i)\,\mathbf{e}_i+a^i\,d\mathbf{e}_i=\mathbf{0} \tag{3}$$

By the formula for the total differential, the quantities $d\mathbf{e}_i$ are expressed in the form

$$d\mathbf{e}_i=\frac{\partial \mathbf{e}_i}{\partial x^j}dx^j=\Gamma_{ij}^k\,\mathbf{e}_k\,dx^j \tag{4}$$

where Γ_{ij}^k are the coefficients of resolution of the vector $\frac{\partial \mathbf{e}_i}{\partial x^j}=\frac{\partial^2 \mathbf{r}}{\partial x^i\,\partial x^j}$ with respect to the basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. These important coefficients were introduced in 1869 by E.B. Christoffel (1829-1900), a German mathematician, and are called the **Christoffel symbols of the second kind**. They make it possible to establish the connection between equal vectors applied in different points in space and are sometimes also denoted by $\left\{ \begin{matrix} i & j \\ k & \end{matrix} \right\}$. The indices in these symbols are placed in such a way that it is possible to perform the necessary summations and apply the summation convention (see Sec. 1.6), but it should be stressed that *the Christoffel symbols are not tensors* because, at each point in space, they not only depend on the basis chosen at that point but also on the variation of the basis as the variable point moves in space. Therefore, the transformation of the Christoffel symbols, when the basis is changed, is performed according to a more complicated rule than in the case of tensors, but we shall not dwell on this question here. We also note that $\Gamma_{ij}^k=\Gamma_{ji}^k$ and that for any affine coordinate system we always have $\Gamma_{ij}^k=0$ (why?).

Substituting (4) into (3) and equating the coefficients in the same base vectors we arrive at the relations

$$da^k+a^i\,\Gamma_{ij}^k\,dx^j=0 \qquad (k=1, 2, \ldots, n) \tag{5}$$

whose collection is equivalent to the condition $da = 0$ and thus expresses the solution of the problem we have set.

If (R) is not only a linear but also a Euclidean space the Christoffel symbols can be expressed in terms of the components of the metric tensor (Sec. 1.7). To this end, we multiply scalarly the equality $\frac{\partial \mathbf{e}_i}{\partial x_j} = \Gamma_{ij}^k \mathbf{e}_k$ by $\mathbf{e}_l$, which yields (see (1.19)) the relation

$$\frac{\partial \mathbf{e}_i}{\partial x_j} \cdot \mathbf{e}_l = g_{lk} \Gamma_{ij}^k \tag{6}$$

The expression on the right-hand side of (6) is called the **Christoffel symbol of the first kind** and is denoted by $\Gamma_{l,\,ij}$, $[ij,\ l]$ or $\left[\begin{smallmatrix} i \\ l \end{smallmatrix} \begin{smallmatrix} j \end{smallmatrix}\right]$. Thus we have

$$\Gamma_{l,\,ij} = g_{lk} \Gamma_{ij}^k, \quad \Gamma_{ij}^k = g^{lk} \Gamma_{l,\,ij} \tag{7}$$

the last formula being implied by the fact that the matrices (g_{lk}) and (g^{lk}) are mutually inverse (see Sec. 1.7). Now let us differentiate the equality $\mathbf{e}_l \cdot \mathbf{e}_k = g_{lk}$ with respect to x^m, then, by virtue of formulas (4) and (6), we obtain

$$\Gamma_{k,\,lm} + \Gamma_{l,\,km} = \frac{\partial g_{lk}}{\partial x^m} \tag{8}$$

Performing twice the cyclic permutation of the indices and taking into account that we always have $\Gamma_{r,\,st} = \Gamma_{r,\,ts}$ we obtain two relations which, together with (8), form a system of three equations with three unknowns. Solving this system we obtain (check it up!)

$$\Gamma_{l,\,km} = \frac{1}{2}\left(\frac{\partial g_{lk}}{\partial x^m} + \frac{\partial g_{lm}}{\partial x^k} - \frac{\partial g_{km}}{\partial x^l}\right)$$

From the last relation we derive, by means of (7), the formula

$$\Gamma_{ij}^k = \frac{1}{2} g^{lk}\left(\frac{\partial g_{li}}{\partial x^j} + \frac{\partial g_{lj}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^l}\right) \tag{9}$$

Consider an example. Let us compute the Christoffel symbols of the second kind for the polar coordinate system ρ, φ in the plane. We have

$$\mathbf{r} = \rho \cos \varphi \mathbf{i} + \rho \sin \varphi \mathbf{j}$$

and therefore, replacing the indices 1 and 2 by ρ and φ, we receive

$$\mathbf{e}_\rho = \mathbf{r}'_\rho = \cos \varphi \mathbf{i} + \sin \varphi \mathbf{j}, \quad \mathbf{e}_\varphi = \mathbf{r}'_\varphi = -\rho \sin \varphi \mathbf{i} + \rho \cos \varphi \mathbf{j} \tag{10}$$

(Note that the notation $\mathbf{e}_\rho$, $\mathbf{e}_\varphi$ is sometimes used for the normalized vectors having unit lengths and going in the same directions;

e.g. see [107], Sec. XVI.28.) It follows that

$$\frac{\partial \mathbf{e}_\rho}{\partial \rho}=0,\ \frac{\partial \mathbf{e}_\rho}{\partial \varphi}=\frac{\partial \mathbf{e}_\varphi}{\partial \rho}=-\sin\varphi\mathbf{i}+\cos\varphi\mathbf{j}$$

$$\frac{\partial \mathbf{e}_\varphi}{\partial \varphi}=-\rho\cos\varphi\mathbf{i}-\rho\sin\varphi\mathbf{j}$$

Expanding these vectors with respect to basis (10) we find the required coefficients $\Gamma^\rho_{\rho\rho}=\Gamma^\varphi_{\rho\rho}=\Gamma^\rho_{\rho\varphi}=\Gamma^\varphi_{\varphi\varphi}=0,\ \Gamma^\varphi_{\rho\varphi}=\frac{1}{\rho},\ \Gamma^\rho_{\varphi\varphi}=-\rho$ (check up the result!).

If the coordinate system x_i is orthogonal the components of the covariant metric tensor relative to the corresponding basis $\mathbf{e}_i$ are $g_{ii}=l_i^2$ and $g_{ij}=0$ $(i\neq j)$ where l_i are *Lamé's coefficients*. Consequently, in this case the symbols Γ^k_{ij} can be easily expressed in terms of Lamé's coefficients with the aid of formula (9). (Let the reader apply this method for computing the Christoffel symbols of the second kind for the polar coordinates.)

In conclusion, we shall derive the differentiation formulas for the reciprocal basis (Secs. 1.6 and 7). For the sake of simplicity, we shall take a Euclidean space (R) although the same result is valid in the case of a general linear space. Let us differentiate equality (1.18) with respect to x^k and substitute, into the resultant relation, the expansions of the derivatives of the base vector with respect to the corresponding bases. This results in

$$0=\frac{\partial \mathbf{e}_i}{\partial x^k}\cdot\mathbf{e}^j+\mathbf{e}_i\cdot\frac{\partial \mathbf{e}^j}{\partial x^k}=\Gamma^s_{ik}\mathbf{e}_s\cdot\mathbf{e}^j+\mathbf{e}_i\cdot M_{jks}\mathbf{e}^s$$

where M_{jks} are some coefficients. Using equalities (1.18) we now obtain $\Gamma^j_{ik}+M_{jki}=0$, that is $M_{jki}=-\Gamma^j_{ik}$. Ultimately, changing the notation of the indices we arrive at the formulas

$$\frac{\partial \mathbf{e}_i}{\partial x^j}=\Gamma^k_{ij}\mathbf{e}_k,\quad \frac{\partial \mathbf{e}^i}{\partial x^j}=-\Gamma^i_{jk}\mathbf{e}^k \tag{11}$$

3. Covariant Differentiation. Consider a Euclidean space with curvilinear coordinates x^i and a vector field $\mathbf{a}=a^i\mathbf{e}_i$ defined in it. Then, by virtue of Sec. 2, we obtain (check it up!)

$$d\mathbf{a}=(da^i+a^j\Gamma^i_{jk}dx^k)\,\mathbf{e}_i=\left(\frac{\partial a^i}{\partial x^k}+\Gamma^i_{jk}a^j\right)\mathbf{e}_i\mathbf{e}^k\cdot dx^s\,\mathbf{e}_s \tag{12}$$

The vector $dx^s\,\mathbf{e}_s=d\mathbf{r}$ being taken quite arbitrarily and, like $d\mathbf{a}$, being independent of the choice of the coordinate system (in particular, of the basis $\mathbf{e}_i$), the expression

$$\nabla\mathbf{a}=\left(\frac{\partial a^i}{\partial x^k}+\Gamma^i_{jk}a^j\right)\mathbf{e}_i\mathbf{e}^k \tag{13}$$

is a (mixed) tensor of rank two independent of the particular choice of the coordinate system. This tensor is called the **cova-**

riant derivative of the vector (vector field) **a**, and its components are denoted as

$$a^i_{,\,k} = \frac{\partial a^i}{\partial x^k} + \Gamma^i_{jk} a^j$$

The derivative of a tensor of any rank is defined similarly. For instance, let the reader use formulas (11) to verify that

$$\nabla (a^{ij}_{\cdot\cdot k} \mathbf{e}^i \mathbf{e}_j \mathbf{e}^k) =$$

$$= \left(\frac{\partial a^{ij}_{\cdot\cdot k}}{\partial x^l} \mathbf{e}^i \mathbf{e}_j \mathbf{e}^k + a^{ij}_{\cdot\cdot k} \frac{\partial \mathbf{e}^i}{\partial x^l} \mathbf{e}_j \mathbf{e}_k + a^{ij}_{\cdot\cdot k} \mathbf{e}^i \frac{\partial \mathbf{e}_j}{\partial x^l} \mathbf{e}^k + a^{ij}_{\cdot\cdot k} \mathbf{e}^i \mathbf{e}_j \frac{\partial \mathbf{e}^k}{\partial x^l} \right) \mathbf{e}^l =$$

$$= \left(\frac{\partial a^{ij}_{\cdot\cdot k}}{\partial x^l} - \Gamma^s_{li} a^{sj}_{\cdot\cdot k} + \Gamma^j_{sl} a^{is}_{\cdot\cdot k} - \Gamma^s_{lk} a^{ij}_{\cdot\cdot s} \right) \mathbf{e}^i \mathbf{e}_j \mathbf{e}^k \mathbf{e}^l$$

(here, when combining similar terms, we change the notation of the dummy indices: $a^{ij}_{\cdot\cdot k} \Gamma^i_{ls} \mathbf{e}^s = a^{sj}_{\cdot\cdot k} \Gamma^s_{li} \mathbf{e}^i$, etc.). It follows that

$$a^{ij}_{\cdot\cdot k,\,l} = \frac{\partial a^{ij}_{\cdot\cdot k}}{\partial x^l} - \Gamma^s_{li} a^{sj}_{\cdot\cdot k} + \Gamma^j_{sl} a^{is}_{\cdot\cdot k} - \Gamma^s_{kl} a^{ij}_{\cdot\cdot s} \qquad (14)$$

Thus, the rank of the resultant tensor is higher by unity than that of the original tensor, the additional index being covariant. (Let the reader carefully examine the structure of the right-hand side of (14).)

In particular, the covariant derivative of a scalar field u is equal to

$$\nabla u = \frac{\partial u}{\partial x^l} \mathbf{e}^l = g^{ij} \frac{\partial u}{\partial x^j} \mathbf{e}_i$$

and thus is the *gradient* of the field. Contracting field (13) we obtain the *divergence* of the vector field:

$$\nabla \cdot \mathbf{a} = \operatorname{div} \mathbf{a} = a^i_{,\,i} = \frac{\partial a^i}{\partial x^i} + \Gamma^i_{ji} a^j \qquad (15)$$

This expression can be given another form if we use the determinant $g = \det (g_{ij})$ and take into account that the derivative of a determinant with respect to its element is always equal to the cofactor of that element (why?). In fact, these cofactors are connected in the well-known way with the elements of the inverse matrix and therefore, applying formulas (7) and (8), we derive the relation

$$\frac{\partial g}{\partial x^l} = \frac{\partial g}{\partial g_{jk}} \frac{\partial g_{jk}}{\partial x^l} = g g^{jk} \frac{\partial g_{jk}}{\partial x^l} = g g^{jk} (\Gamma_{k,\,jl} + \Gamma_{j,\,kl}) = g (\Gamma^j_{jl} + \Gamma^k_{kl}) = 2g\Gamma^j_{jl}$$

Consequently, the right-hand side of (15) can be rewritten in the form

$$\operatorname{div} \mathbf{a} = \frac{\partial a^i}{\partial x^i} + \frac{1}{2g} \frac{\partial g}{\partial x^l} a^i = \frac{1}{\sqrt{g}} \frac{\partial (\sqrt{g}\, a^i)}{\partial x^i}$$

It is readily proved that the rotation of a vector field in the three-dimensional space is expressed by the formula

$$\operatorname{rot}\mathbf{a}=\frac{1}{\sqrt{g}}\left(\left(\frac{\partial a_3}{\partial x^2}-\frac{\partial a_2}{\partial x^3}\right)\mathbf{e}_1+\left(\frac{\partial a_1}{\partial x^3}-\frac{\partial a_3}{\partial x^1}\right)\mathbf{e}_2+\left(\frac{\partial a_2}{\partial x^1}-\frac{\partial a_1}{\partial x^2}\right)\mathbf{e}_3\right) \quad (16)$$

where $a_i=g_{ij}a^j$ (see Sec. 1.7). To obtain this formula (here we do not present its deduction) one must show that the right-hand side of (16) is a vector, i. e. is invariant under any transformation of coordinates, and then conclude that formula (16) is valid for any coordinate system since it is obviously true for Cartesian coordinates.

Formula (12) is readily generalized to tensors of any rank. It follows that

$$d\mathbf{T}=\nabla\mathbf{T}\cdot d\mathbf{r}=(\nabla_i\mathbf{T})\,dx^i \quad (17)$$

where the inner product is contracted (see Sec. 1.3) with respect to the covariant index of $\nabla_i\mathbf{T}$ and contravariant index of dx^i; the expressions $\nabla_i\mathbf{T}$ are understood as the components of the tensor $\nabla\mathbf{T}$ which are determined by the formula $\nabla\mathbf{T}=(\nabla_i\mathbf{T})\,\mathbf{e}^i$. In particular, relation (17) implies the formula

$$\frac{d\mathbf{T}}{dt}=\frac{\partial\mathbf{T}}{\partial t}+\nabla\mathbf{T}\cdot\mathbf{v}\quad\left(\mathbf{v}=\frac{d\mathbf{r}}{dt}=\frac{dx^i}{dt}\,\mathbf{e}_i\right)$$

expressing the rate of change of an arbitrary tensor field $\mathbf{T}$ along a trajectory of the velocity field $\mathbf{v}$ (cf. Sec. I.1.1). In particular, for a scalar field u and a vector field $\mathbf{a}$ we obtain

$$\frac{du}{dt}=\frac{\partial u}{\partial t}+\operatorname{grad}u\cdot\mathbf{v}=\frac{\partial u}{\partial t}+\frac{\partial u}{\partial x^i}\frac{dx^i}{dt}$$

and

$$\frac{d\mathbf{a}}{dt}=\frac{\partial\mathbf{a}}{\partial t}+\nabla\mathbf{a}\cdot\mathbf{v}=\left(\frac{\partial a^i}{\partial t}+\left(\frac{\partial a^i}{\partial x^k}+\Gamma^i_{jk}a^j\right)v^k\right)\mathbf{e}_i$$

The ordinary rules for differentiating sums and products are readily extended to the covariant derivative. Indeed, the latter is expressed in terms of the differential for which these rules hold.

The covariant derivative of a metric tensor $\mathbf{g}=g_{ij}\mathbf{e}^i\mathbf{e}^j$ *is always equal to zero.* Virtually, in Cartesian coordinates we have $\mathbf{g}=\mathbf{e}_1\mathbf{e}_1+\ldots+\mathbf{e}_n\mathbf{e}_n$, that is this tensor is constant in the whole space; consequently $d\mathbf{g}=\mathbf{0}$, and therefore $\nabla\mathbf{g}=\mathbf{0}$ (why?). Analogously, $\nabla(g^{ij}\mathbf{e}_i\mathbf{e}_j)=\mathbf{0}$. In other words, we can write $g_{ij,\,l}=g^{ij}_{\ \ ,l}=0$.

A field can be differentiated repeatedly. In particular, an important role is played by the operator Δ_2 determined by the formula

$$\Delta_2 u=\operatorname{div}\operatorname{grad}u=\frac{1}{\sqrt{g}}\frac{\partial}{\partial x^i}\left(\sqrt{g}\;g^{ij}\frac{\partial u}{\partial x^j}\right) \quad (18)$$

which is known as the **Laplace-Beltrami operator** (E. Beltrami, 1835-1900, an Italian geometer). This operator is independent

of the choice of the coordinate system and turns into the Laplacian operator (see (I.1.7)) in every Cartesian system of coordinates. We thus can regard (18) as an expression of the Laplacian in an arbitrary coordinate system.

We can similarly obtain an expression, in arbitrary curvilinear coordinates, for any differential operator applied to a given field. To this end, it is sufficient to write the expression in question in the tensor (invariant) notation by applying the rules of tensor calculus, which results in a formula taking a given concrete form in a concrete coordinate system and invariant with respect to the coordinate transformations. The tensor form of basic equations is widely used in theoretical physics, mechanics of continua and the like.

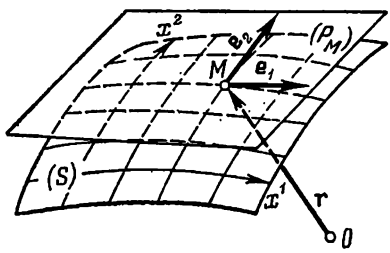


Fig. 74

4. Tensor Field on a Manifold of a Euclidean Space. To begin with, let us consider an ordinary two-dimensional surface (S) in the ordinary three-dimensional geometric space. Suppose that there is a system of general curvilinear coordinates x^1, x^2 defined on (S). Then, at each point $M \in (S)$, the vectors $\mathbf{e}_1 = \frac{\partial \mathbf{r}}{\partial x^1}$ and $\mathbf{e}_2 = \frac{\partial \mathbf{r}}{\partial x^2}$ are in the tangent plane (P_M) to (S) drawn through M (see Fig. 74). When the coordinate system is transformed these vectors are changed according to the same rule as that given at the beginning of Sec. 2; this makes it possible to define scalar, vector and, generally, tensor fields on (S) in just the same way as in an arbitrary Euclidean space.

For small variations of the coordinates we have

$$d\mathbf{r} = \frac{\partial \mathbf{r}}{\partial x^1} dx^1 + \frac{\partial \mathbf{r}}{\partial x^2} dx^2 = (dx^i)\, \mathbf{e}_i$$

and the square of length of the vector $d\mathbf{r}$ is equal to

$$ds^2 = d\mathbf{r} \cdot d\mathbf{r} = (dx^i \mathbf{e}_i) \cdot (dx^j \mathbf{e}_j) = (\mathbf{e}_i \cdot \mathbf{e}_j)\, dx^i\, dx^j = g_{ij}\, dx^i\, dx^j \qquad (19)$$

This expression is called the **first fundamental quadratic (differential) form of the surface** (S) (it was introduced in 1828 by K.F. Gauss). The coefficients g_{ij} of the quadratic form, known as the **fundamental coefficients of** (S) or, more precisely, the **fundamental coefficients of the first order of** (S), constitute the **covariant metric tensor of the surface** (S). If form (19) is given, we can compute the angle between two lines on (S) at the point

of their intersection by means of the formula

$$\cos(d\mathbf{r}, d\tilde{\mathbf{r}}) = \frac{d\mathbf{r}\cdot d\tilde{\mathbf{r}}}{ds\,d\tilde{s}} = \frac{g_{ij}\,dx^i\,d\tilde{x}^j}{\sqrt{g_{ij}dx^i dx^j}\sqrt{g_{ij}d\tilde{x}^i d\tilde{x}^j}}$$

The arc length of any curve (L) lying on (S) is found with the aid of the formula

$$L = \int\limits_{(L)} ds = \int\limits_{(L)} \sqrt{g_{ij}(x^1, x^2)\,dx^i\,dx^j}$$

and the area of any figure (σ) on (S) is expressed by the formula

$$\begin{aligned}\sigma &= \int\limits_{(\sigma)} |\partial_{x^1}\mathbf{r} \times \partial_{x^2}\mathbf{r}| = \\ &= \int\limits_{(\sigma)} \sqrt{|\mathbf{e}_1 \times \mathbf{e}_2|^2}\,dx^1 dx^2 = \\ &= \int\limits_{(\sigma)} \sqrt{(|\mathbf{e}_1|^2|\mathbf{e}_2|^2 - (\mathbf{e}_1\cdot\mathbf{e}_2)^2)^2}\,dx^1\,dx^2 = \\ &= \int\limits_{(\sigma)} \sqrt{g}\,dx^1\,dx^2 \qquad (g = \det(g_{ij}))\end{aligned} \tag{20}$$

(Let the reader carefully examine these formulas.)

Formulas of type (11) are no longer valid in this case and should be appropriately changed since, generally speaking, a vector of the form $\frac{\partial \mathbf{e}_i}{\partial x^j}$ is not necessarily parallel to the plane (P_M) (why?). Therefore, we shall limit ourselves to expanding the projections of this vector on the plane (P_M) with respect to the vectors $\mathbf{e}_i$:

$$\mathrm{proj}_{(P)} \frac{\partial \mathbf{e}_i}{\partial x^j} = \Gamma_{ij}^k \mathbf{e}_k \tag{21}$$

(in this formula we do not write the index M because expansion (21) applies to any point of the surface (S)). The expression $\Gamma_{ij}^k \mathbf{e}_k$ (which we have called the projection of $\frac{\partial \mathbf{e}_i}{\partial x^j}$ on (P_M)) is the component (lying in (P_M)) of the vector $\frac{\partial \mathbf{e}_i}{\partial x^j}$ when it is resolved along the direction normal to (P_M) at the point M and along the corresponding direction in the tangent plane (P_M). For the coefficients Γ_{ij}^k, understood in this sense, formula (6) remains valid and also implies the validity of formula (9) expressing the coefficients Γ_{ij}^k (check it up!). Both formulas (11) also remain true in this sense (i.e. if their right-hand and left-hand sides are understood as the corresponding projections on the tangent plane).

Consider as an example a sphere of radius R. Let us introduce spherical coordinates θ, φ on it (e.g. see [107], Sec. X.1). Here we have

$$\mathbf{r} = R(\sin\theta\cos\varphi\mathbf{i} + \sin\theta\sin\varphi\mathbf{j} + \cos\theta\mathbf{k})$$

$$\mathbf{e}_\theta = \frac{\partial\mathbf{r}}{\partial\theta} = R(\cos\theta\cos\varphi\mathbf{i} + \cos\theta\sin\varphi\mathbf{j} - \sin\theta\mathbf{k})$$

$$\mathbf{e}_\varphi = R(-\sin\theta\sin\varphi\mathbf{i} + \sin\theta\cos\varphi\mathbf{j})$$

whence

$$g_{\theta\theta} = \mathbf{e}_\theta \cdot \mathbf{e}_\theta = R^2, \quad g_{\theta\varphi} = 0, \quad g_{\varphi\varphi} = R^2\sin^2\theta$$

The application of formula (9) yields the expressions

$$\Gamma^\theta_{\varphi\varphi} = \sin\theta\cos\theta, \quad \Gamma^\varphi_{\theta\varphi} = \cot\theta, \quad \Gamma^\theta_{\theta\theta} = \Gamma^\theta_{\theta\varphi} = \Gamma^\varphi_{\theta\theta} = \Gamma^\varphi_{\varphi\varphi} = 0$$

(Let the reader also derive these formulas by means of expansion (21).)

In the case of an arbitrary surface (S) the considerations given at the beginning of Sec. 3 should also be appropriately changed because, generally speaking, the vector $d\mathbf{a}$ has an additional component directed along the normal to (S). But the definition of the covariant derivative is in this case formulated in terms of the projections on the plane (P_M) (i.e. after the above component has been discarded), and therefore formula (13) nevertheless remains valid, the tensor $\nabla\mathbf{a}$ being independent of the choice of the coordinate system on (S).

Applying the same operation of projecting on the tangent plane we can generalize the notion of Christoffel's symbols to an arbitrary n-dimensional manifold lying in an m-dimensional space and also the notion of the covariant derivative of a tensor field of any rank on this manifold. The basic formulas of Secs. 2 and 3 related to these notions remain valid in this case but the summands involving total differentials may additionally contain some terms corresponding to normal components. The proposition asserting that $\nabla\mathbf{g} = \mathbf{0}$ also remains true although the proof given in Sec. 3 no longer applies. But this proposition can be formally proved by calculating $g_{ij,l}$ according to formulas analogous to (14) and substituting expressions of type (9) for the Christoffel symbols into them, and hence, since these formulas and expressions apply to an arbitrary manifold as well, the equality $g_{ij,l} = 0$ implied by them is always fulfilled.

5. Intrinsic Geometry and Riemannian Spaces. Let us come back to the beginning of Sec. 4 and suppose that the surface (S) undergoes bending without stretching and shrinking as if it were a flexible inextensible film on which the net of coordinate lines is marked. Then all arc lengths on (S) remain unchanged, and therefore the quadratic form (19) is invariant and all its coefficients retain their values although the functional relationship $\mathbf{r} = \mathbf{r}(x_1, x_2)$

is changed. Conversely, if there are two surfaces (S) and $(\tilde{S})$ with coordinate systems x^i and $\tilde{x}^i$ on them, and if there is a one-to-one mapping of (S) onto $(\tilde{S})$ such that $x^i = \tilde{x}^i$ for the corresponding points and quadratic forms (19) coincide (i.e. $g_{ij} = \tilde{g}_{ij}$), these surfaces can be applied to each other without stretching and shrinking in such a way that the corresponding points coincide. In fact, according to Sec. 4, the lengths, angles and areas are invariant under this mapping, because they are expressible in terms of the coefficients g_{ij}. Two such surfaces (S) and $(\tilde{S})$ are said to be **isometric or applicable**.

All geometric properties that are the same for isometric surfaces are called the **intrinsic** or **absolute** properties and constitute the **intrinsic geometry** of these surfaces. For instance, small portions of a plane, cylindrical and conic surfaces have the same intrinsic geometry because the latter two can be developed (rolled out) on the plane. All the quantities pertaining to a surface which are expressible in terms of the coefficients g_{ij} (these are not only lengths, angles and areas but also the Christoffel symbols and the quantities resulting from differentiation of fields defined on the surface which are expressible in terms of g_{ij}) are included into the intrinsic geometry of the surface. On the other hand, the curvature of a line on the surface or the shortest distance between two points of the surface (understood in the ordinary sense as the length of the line segment joining these points in space) does not belong to the intrinsic geometry because these quantities vary as the surface is being bent. The total differential $d\mathbf{a}$ of the field of the tangent vectors $\mathbf{a}$ defined on the surface does not belong to the intrinsic geometry either because, under bending, the normal component of $d\mathbf{a}$ is changed (cf. Sec. 4).

The intrinsic geometry of an arbitrary n-dimensional manifold is defined in an analogous manner. From the point of view of the intrinsic geometry, it is inessential in what way the manifold is embedded in the enveloping Euclidean space of higher dimension, and the only important fact is the dependence of its metric tensor on the coordinates on the manifold.

To pass to the general notion of a *Riemannian space*, it now only remains to discard the notion of an enveloping Euclidean space and consider only the manifold itself and the metric tensor defined on it. This step was made by G.F.B. Riemann in 1854; it is very important because a manifold understood as a collection of some objects with n degrees of freedom may appear without any connection with a Euclidean space (e.g. see [107], Sec. X.2).

Thus, a **Riemannian space** (R) is an n-dimensional manifold at whose every point a metric tensor g_{ij} is defined. The latter is a covariant tensor of order two, that is the coefficients g_{ij} assume

at each point of the space (R) some concrete values, after a coordinate system x^i has been introduced in a neighbourhood of the point, and are transformed according to the tensor rule

$$g'_{ij}=\frac{\partial x^{i_1}}{\partial x^{i'}}\frac{\partial x^{j_1}}{\partial x^{j'}}g_{i_1j_1}$$

when we pass to another coordinate system $x^{i'}$. The notion of an element of length is introduced by means of the metric tensor:

$$ds^2=g_{ij}dx^idx^j \tag{22}$$

For this element to be positive it is required that quadratic form (22) be positive definite.

As in Sec. 4, knowing the expression of the element of length we can determine the length of any curve in (R) and the angle between two curves at their intersection point. Instead of formula (20), we obtain the formula $\int\limits_{(\sigma)}\sqrt{g}\,dx^1dx^2\ldots dx^n=\sigma$ for the n-dimensional volume of a domain (σ) because it can be shown that this formula determines the n-dimensional volume of an n-dimensional domain in a Euclidean space. Let the reader deduce the formula for computing the k-dimensional volume $(2\leqslant k\leqslant n-1)$ of a domain of a k-dimensional manifold lying in (R) by taking into account that such a manifold is itself a Riemannian space. Knowing the formulas for computing volumes we can introduce, in the ordinary way, the notion of an integral of a function over a domain of (R) or over a domain belonging to a manifold of dimension $k<n$ lying in (R).

The introduction of the notion of a vector field in (R) is connected with a difficulty because, generally speaking, (R) is not a linear space. To understand the nature of this difficulty imagine that we are trying to study vectors on a curvilinear surface without considering the space enveloping this surface. But this difficulty is not essential and can be overcome. The matter is that for constructing the field theory it is sufficient to use only infinitesimal vectors and consider every finite vector not geometrically but as the result of a normalization of the corresponding infinitesimal vector by multiplying the latter by a scalar factor. For example, the velocity vector $\mathbf{v}=\frac{d\mathbf{r}}{dt}$ can be represented not by a vector of finite length but in the form of the infinitesimal vector $d\mathbf{r}$ supplied with the normalization factor $\frac{1}{dt}$ or, which is (formally) even simpler, as a small vector of indefinite length going in the direction of $d\mathbf{r}$ with the additional indication of its modulus, i.e. the scalar field $v=|\mathbf{v}|$. Following this idea, we can

determine a vector field $\mathbf{a}$ on (R) by indicating, at each point $M \in (R)$, the direction of the field with the aid of an infinitesimal vector and the value of the modulus a of $\mathbf{a}$ attributed to that infinitesimal vector. It turns out that this makes it possible to perform any operations on these vectors which are thought of as starting from the corresponding points M although they are not actually laid off from these points. In particular, if there is a coordinate system x^i defined on (R) the vector $\mathbf{e}_1$ (and, similarly, the vectors $\mathbf{e}_2, \ldots, \mathbf{e}_n$) is defined at each point M by specifying the direction of the vector $\partial_{x^1}\mathbf{r}$ (while the vector $\mathbf{r}$ itself is not considered) and setting the modulus equal to $\left|\frac{\partial \mathbf{r}}{\partial_{x^1}}\right| = \sqrt{g_{11}}$. This enables us to write vector and tensor fields in the ordinary form $\mathbf{a} = a^i\mathbf{e}_i$, $\mathbf{g} = g_{ij}\mathbf{e}^i\mathbf{e}^j$ and the like and perform on them algebraic operations which reduce to operations on vectors or tensors taken at one and the same point.

Now let us discuss a more complicated question concerning the rules of translation of a vector in (R). Without these rules it is impossible to operate on vectors defined in different points of (R) and, in particular, to introduce the expression $d\mathbf{a}$ (why?). We must formulate the rules in such a way that, for a manifold lying in a Euclidean space, they reduce to a procedure expressed in terms of the intrinsic geometry of the manifold. We have already mentioned that for such a manifold the differential $d\mathbf{a}$ includes two summands, namely the tangential term $\nabla\mathbf{a}\cdot\mathbf{dr}$ invariant under the bending of the manifold and the normal term changed under the bending. Therefore, it appears natural to define $d\mathbf{a}$ by means of the formula

$$d\mathbf{a} = \nabla\mathbf{a}\cdot\mathbf{dr} \tag{23}$$

where $\mathbf{dr}$ is an infinitesimal displacement vector and $\nabla\mathbf{a}$, as before, is computed according to formula (13) in which the coefficients Γ^i_{jk} are expressed by means of formula (9) in terms of the elements of the metric tensor defined on (R). The differential of a tensor field of an arbitrary rank is defined in like manner.

The differential and covariant derivatives of the fields in a Riemannian space possess properties similar to those described in Secs. 2 and 3 for the fields defined in a Euclidean space, but here there are also some new features related to the translation of a vector $\mathbf{a}$ when its origin is moved along a curve. When this procedure is performed we must have $d\mathbf{a} = \mathbf{0}$, and therefore, as in Sec. 2, relations (5) should hold. It turns out that in the general case, in every (even small but finite) neighbourhood of the initial point, the direction of the translated vector depends on the path. Therefore, generally speaking, there are no domains in a Riemannian space in which there exist fields of parallel vectors.

To visualize this situation, let us consider the example of a two-dimensional sphere regarded as a two-dimensional Riemannian space. This means that we only take into account the properties of the sphere related to its intrinsic geometry. The vectors of this space can be represented as the ones tangent to the sphere, and the translation of such a vector should be understood in the sense of the intrinsic geometry of this Riemannian space, which means that for every infinitesimal displacement of the origin of the vector its projection on the tangent plane must gain an increment which is an infinitesimal of higher order relative to the displacement. Imagine that the vector is moved in such a way that its origin traverses the contour $ABCA$ (see Fig. 75) which is composed of an arc of the equator and two meridional arcs. If the vector $\mathbf{a}$ occupies the position $\mathbf{a}_0$ at the initial point A, then in the process of translation it passes through the positions depicted in Fig. 75 (this is also implied by the symmetry considerations). We see that, after the closed contour has been described, the translated vector comes to the position $\mathbf{a}_1$ distinct from $\mathbf{a}_0$.

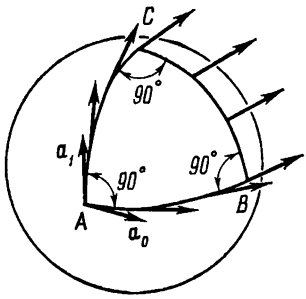


Fig. 75

If there are two vectors which are translated simultaneously their scalar product retains its value since we have $d(\mathbf{a}\cdot\mathbf{b}) = d\mathbf{a}\cdot\mathbf{b} + \mathbf{a}\cdot d\mathbf{b} = 0$. (It should be noted that the condition $d\mathbf{a} = \mathbf{0}$ no longer implies that $\mathbf{a} = \text{const}$ because, in a general Riemannian space, the latter relation is senseless.) In particular, it follows that the modulus of a vector and the angle between two vectors are not changed under translation (why?).

Suppose that in a two-dimensional Riemannian space (R) we choose a domain (σ) with bounding contour (L). Let $\alpha_{(\sigma)}$ be the angle through which an arbitrary vector $\mathbf{a}$ turns when it is translated in such a way that its origin moves along (L) (the angle is considered positive if, as in Fig. 75, the vector turns in the direction of traversing the contour). It is clear, from the foregoing paragraph, that the angle $\alpha_{(\sigma)}$ is independent of the concrete choice of the vector $\mathbf{a}$. It can also be readily shown that it is independent of the choice of the initial point on (L) either. (Strictly speaking, this definition only applies to sufficiently small domains (σ) but this is sufficient for our aims; if necessary it is possible to pass to arbitrary domains by deforming the small ones in a continuous manner.)

It can also be proved (how?) that the angle $\alpha_{(\sigma)}$ satisfies the law of addition, namely, if the domain (σ) is divided into two portions (σ_1) and (σ_2) then $\alpha_{(\sigma)} = \alpha_{(\sigma_1)} + \alpha_{(\sigma_2)}$. This makes it pos-

sible to apply the standard technique (e.g. see [107], Sec. XVI.7) in order to introduce the density of the quantity α:

$$K(M) = \lim_{(\sigma)\to M} \frac{\alpha_{(\sigma)}}{\sigma} \qquad (M \in (R)) \tag{24}$$

This density is referred to as the *curvature* of the space (R) at the point M. Conversely, the quantity α is equal to the integral of its density:

$$\alpha_{(\sigma)} = \int_{(\sigma)} K \, d\sigma$$

For instance, for the sphere shown in Fig. 75 the ratio $\frac{\alpha_{(\sigma)}}{\sigma}$ corresponding to the domain $ABCA$ is equal to $\frac{\left(\frac{\pi}{2}\right)}{\left(\frac{4\pi R^2}{8}\right)} = \frac{1}{R^2}$ where R is the radius of the sphere (check it up!). But the sphere possesses the same properties at all its points, and consequently its curvature is one and the same everywhere and thus is equal to $\frac{1}{R^2}$. It is possible to prove that for every two-dimensional surface embedded in the three-dimensional space (but regarded as a Riemannian space) the curvature introduced here in terms of the intrinsic geometry coincides with the *total (Gaussian) curvature* (on the notion of the total curvature see, for instance, [107], Sec. XII.9).

In a Riemannian space (R) of dimension equal to or exceeding 3, limit (24) is not only dependent on the point M but also on the orientation of the surface element (σ) in space. Therefore, the curvature of (R) at M is no longer a scalar. It can be proved that this curvature is completely characterized by a tensor referred to as the **curvature tensor** which is of rank 4 and is expressed in terms of the Christoffel symbols of the second kind of the space (R).

In conclusion we note that in recent years much attention has been paid to the theory of manifolds with metric tensors defined on them for which quadratic forms (22) are no longer positive definite. These spaces are called **pseudo-Riemannian.** According to the general theory of relativity, the physical four-dimensional space (the so-called **space-time**) is a space of that kind. Its three-dimensional section by every hyperplane $t = \text{const}$, characterizing the state of the physical space at time moment t, is a Riemannian space possessing a curvature tensor at each point. These properties, together with some additional physical hypotheses, imply that the universe is of finite volume (i.e. is closed like a sphere) and make it possible to estimate this volume. It should be noted

that the question "in what direction the universe is curved?" which at first glance appears natural is in fact senseless because the concept of the curvature of the universe can only be approached from the point of view of its intrinsic geometry which does not involve the considerations related to any enveloping space.

A comprehensive representation of the theory of Riemannian spaces can be found in [113] (where the mathematical principles of the relativity theory are also considered) and [31].

CHAPTER VI

Calculus of Variations

The *calculus of variations* was founded in the 18th century by L. Euler and J.L. Lagrange and extensively developed by other mathematicians in the 19th century. At present it is one of the most important divisions of theoretical and applied mathematics. Variational principles are of great scientific significance; they provide general approach to various physical and applied problems and yield fundamental exploratory ideas. Variational methods prove effective in problems solving practice both in qualitative and quantitative aspects.

The calculus of variations has much in common with the theory of extrema of functions (e.g. see [107], §§ IV.6 and XII.2) and is, to a certain extent, a generalization and further development of this theory. Among many courses devoted to the calculus of variations we can recommend [3], [11], [32], [33], [48], [79], [80], [121], [122] and [131].

§ 1. FIRST VARIATION AND NECESSARY CONDITIONS FOR EXTREMUM

1. Examples of Variational Problems.

1. We shall begin with an ancient problem of finding the curve, with a given perimeter, which encloses the maximum area. This is a typical variational problem known as **Dido's problem.** The solution of the problem (which is an arc of a circle) is reported to have been known to Queen Dido of Carthage, who was given as much land as she could enclose with a cowhide. She cut it into strips, made a thin thread and formed a semicircle along the coast of North Africa, within which was established the State of Carthage.

This problem has several modifications. Let us consider the following variant. Suppose that we have a thread of given length whose ends are fixed at the points A and B (see Fig. 76). If we take the coordinate system shown in the figure, the points A and

B have, respectively, the coordinates $(a, 0)$ and $(b, 0)$ and thus we arrive at the problem of maximizing the integral

$$S=\int_a^b y(x)\,dx \tag{1}$$

expressing the enclosed area where $y=y(x)$ is the equation of the sought-for curve. The condition that the length of the thread is

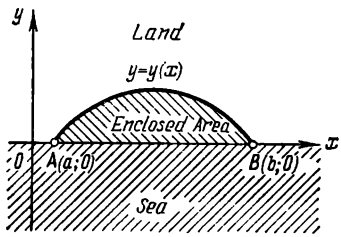


Fig. 76

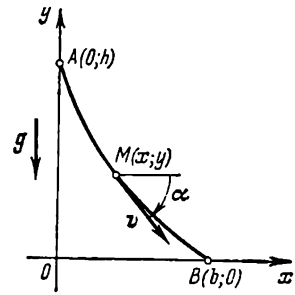


Fig. 77

given and fixed reduces to the additional requirement that the function $y(x)$ satisfy the integral relation

$$L=\int_a^b \sqrt{1+y'^2}\,dx \tag{2}$$

with the given boundary values

$$y(a)=0, \quad y(b)=0 \tag{3}$$

As has been mentioned, the sought-for curve is an arc of a circle (the solution of the problem will be considered in detail in Sec. 11).

2. Another celebrated problem which led to the development of the methods of the calculus of variations was proposed by Johann Bernoulli in 1696 as a challenge to the mathematicians of Europe (the so-called **brachistochrone problem**). This is the problem of finding the equation $y=y(x)$ of the path down which a particle M will fall, under gravitation (without friction), from a given point A to a given point B (Fig. 77) in the shortest time. The problem was investigated with the aid of different methods by Jacob Bernoulli, G.W. Leibniz, G.G.A. L'Hospital and I. Newton. They all found the correct solution which is a cycloid through the two given points. To formulate the problem mathematically we must express the time of motion T in terms of the unknown function $y(x)$. By the law of conservation

of energy we have

$$gm(h-y)=\frac{mv^2}{2}, \text{ i.e. } v=\sqrt{2g(h-y)}$$

whence

$$\frac{dx}{dt}=v\cos\alpha=\frac{\sqrt{2g(h-y)}}{\sqrt{1+y'^2}} \quad \text{and thus} \quad dt=\sqrt{\frac{1+y'^2}{2g(h-y)}}\,dx$$

Therefore, we receive

$$T=\int dt=\int_0^b \sqrt{\frac{1+y'^2}{2g(h-y)}}\,dx \tag{4}$$

Thus, the problem reduces to determining the function $y(x)$ giving the minimum to integral (4). The sought-for function $y(x)$ must satisfy the subsidiary (boundary) conditions

$$y(0)=h,\ y(b)=0 \tag{5}$$

The solution of this problem will be discussed in Sec. 8.

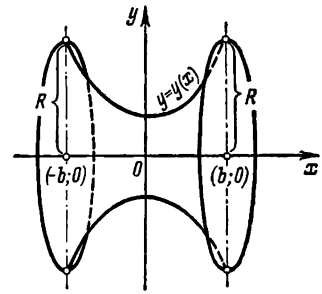


Fig. 78

3. Now we shall consider the problem concerning the equilibrium form of a soap-film spanned by two coaxial wire rings (see Fig. 78). For simplicity, let the rings be congruent, their radius being equal to R. If the film is considered weightless it follows from the theory of surface tension that the sought-for form of the film is of the minimum area *). On the basis of intuitive consideration let us assume that the film will take the form of a surface of revolution. If the equation of the curve whose revolution generates the surface is $y=y(x)$ we arrive at the mathematical problem of minimizing the integral

$$S=2\pi\int_{-b}^{b} y\sqrt{1+y'^2}\,dx \tag{6}$$

which is equal to the area of the surface of revolution. The

*) We thus deal with a particular case of the general problem of finding a surface of minimum area stretched over a given contour. It can be shown that if a smooth surface minimizes the area it is a **minimal surface**, i.e. one whose **mean curvature** (the algebraic sum of the curvatures of its *principal normal section*; e.g. see [107], Sec. XII.2.4) vanishes identically. The problem of determining a minimal surface with a given (twisted) curve as its boundary is called the **Plateau problem** after J.A.F. Plateau, a Belgian physicist, who solved the problem for several different contours in soap-film experiments.—*Tr.*

function $y(x)$ must of course satisfy the boundary conditions

$$y(-b)=R, \qquad y(b)=R \tag{7}$$

Hence, here we have a particular case of the general **minimum-surface-of-revolution problem** of determining a curve with given boundaries whose rotation about an axis generates a surface of minimum area. The solution of the "soap-film" problem will be discussed in Sec. 8.

2. Functional. Now we can discuss some general features of the above problems. First of all, they are extrema problems dealing with maxima or minima of given quantities. An elementary problem of this kind is to find an extremum of a function which is solved on the basis of differential calculus. If we consider a function $f(x)$ of one independent variable x, there is one degree of freedom in the determination of the required value of x giving an extremum to the function (e.g. see [107], § IV.6). If it is necessary to find an extremum of a function $f(x_1, x_2, \ldots, x_n)$ dependent on n variables (e.g. see [107], § XII.2) we look for the values $x_1, x_2, \ldots, x_n$ specifying the point of extremum and thus have n degrees of freedom. But in the problems discussed in Sec. 1 the sought-for objects are curves. This means, analytically, that we must determine a function extremizing a given quantity on condition that it satisfies given boundary conditions. When considering an unknown function we have an infinite number of degrees of freedom (why?), and, consequently, we can say that *the calculus of variations deals with extrema problems involving an infinite number of degrees of freedom.*

Let us consider another characteristic feature of these problems. If we take, for instance, integral (6) and substitute into it an arbitrary function $y(x)$ defined for $-b \leqslant x \leqslant b$ and satisfying boundary conditions (7) this integral assumes a concrete numerical value. When there is a correspondence which attributes to every function, belonging to a given class, a certain numerical value, we say that *there is a* **functional** *defined on that class of functions* (e.g. see [107], Sec. XIV.4; for the sake of simplicity we consider all the quantities involved as being dimensionless and therefore speak about "numerical values"; in the general case we should speak about "scalar quantities"). We thus see that the variational problem we are considering reduces to finding a function $y(x)$ for which integral (6) attains its minimum value, and, consequently, the *calculus of variations studies extrema of functionals.*

Let us consider some examples of concrete functionals. Take the integral

$$I = I\{y\} = \int_0^2 y^2\,dx \tag{8}$$

This is a functional defined for all functions $y(x)$, having the domain of definition $0 \leqslant x \leqslant 2$ and assuming finite (or even infinite) values, such that integral (8) exists (is convergent). For instance, taking the function $y = x^2$ $(0 \leqslant x \leqslant 2)$ we obtain $I = \int_0^2 x^4\,dx = 6.4$. For $y = \sin x$ $(0 \leqslant x \leqslant 2)$ we have $I = 1 - \frac{\sin 4}{4} = 1.19$, for $y = \frac{1}{\sqrt[4]{x}}$ $(0 \leqslant x \leqslant 2)$ we receive $I = 2\sqrt{2} = 2.83$, etc.

It should be stressed that the expression $I\{y\}$ must not be interpreted as a composite function of the type of $F(y(x)) = [y(x)]^2$ and the like. Indeed, for every concrete $y(x)$ the expression $F(y(x))$ is a new function whose value $F(y(x_0))$ is completely determined, for every $x = x_0$, by the particular value $y(x_0)$ of the function $y(x)$. In contrast to it, the functional $I\{y\}$ is not solely determined by certain particular value of the function $y(x)$, because, for every concrete function $y(x)$ the value of I depends on the functional relationship $y = y(x)$ as a whole. We can say that a functional is a function whose argument takes on the values which are ordinary functions $y(x)$ while the values of the functional itself are numbers.

The formula

$$J\{y\} = \int_0^2 (xy + y')\,dx$$

specifies another functional whose important feature is that it is **linear**, that is

$$J\{Cy\} = CJ\{y\} \ (C = \text{const}), \ J\{y_1 + y_2\} = J\{y_1\} + J\{y_2\} \qquad (9)$$

The formula

$$K\{y\} = \int_1^3 y^2\,dx$$

determines a functional different from (8) because the domain of definition of $K\{y\}$ is a class of functions $y(x)$ other than in (8) because they are defined for $1 \leqslant x \leqslant 3$ but not for $0 \leqslant x \leqslant 2$. At the same time, the functional

$$L\{z\} = \int_0^2 z^2\,dx$$

coincides with (8), and we can write $I\{y\} \equiv L\{y\}$.

3. Functional Spaces. When studying a functional and its extrema we must stipulate the **domain of definition** of the functional, that is the class of functions for which the values of the functional are defined. As a rule, a class of that type is a linear

space (e.g. see [107], § VII.6) or its part consisting of functions on which the linear operations are performed according to the well-known rules. These spaces are referred to as **functional spaces**, and they are most often infinite-dimensional.

We usually deal with the so-called **normed linear space** in which the notion of a *norm* is introduced. The **norm** of a function f is denoted as $\|f\|$; it characterizes the deviation of f from the zero function (which is identically equal to zero), is a finite real number and must satisfy the following **axioms of norm**:

1. $\|f\| \geqslant 0$, and $\|f\| = 0$ only for the function $f \equiv 0$.
2. $\|\lambda f\| = |\lambda| \|f\|$ for every $\lambda = \text{const}$.
3. $\|f+g\| \leqslant \|f\| + \|g\|$.

As is known (e.g. see [107], Sec. XVII.7), the deviation of functions can be introduced in various ways. Accordingly, we consider different functional spaces consisting of functions defined on an interval $a \leqslant x \leqslant b$. (We stress that all the functions entering into a linear functional space must be defined on a common interval because, if otherwise, it would be impossible to add them together.) The theory of functional spaces is studied in courses of *functional analysis*. Let us enumerate some spaces which are most often dealt with. (For simplicity, in the present chapter all the quantities involved are regarded as real and all the domains of definition as finite.)

1. **The space of all continuous functions** defined on a finite interval $a \leqslant x \leqslant b$. The norm is determined by the formula

$$\|f\| = \max_{a \leqslant x \leqslant b} |f(x)|$$

and corresponds to the *uniform deviation* of functions. We denote this space by the symbol $C[a, b]$.

2. **The space of all continuously differentiable functions** defined for $a \leqslant x \leqslant b$ (denoted as $C_1[a, b]$). The norm in this space is

$$\|f\| = \max_{a \leqslant x \leqslant b} |f(x)| + \max_{a \leqslant x \leqslant b} |f'(x)| \qquad (10)$$

(As a norm in $C_1[a, b]$ we can also take, instead of sum (10), the greatest of the terms entering into the right-hand side; this results in an inessential distinction between these spaces of continuously differentiable functions.)

We similarly define the space $C_n[a, b]$, $n = 2, 3, 4, \ldots$, consisting of all functions possessing continuous derivatives up to order n inclusive.

3. **The Hilbert space** $L_2[a, b]$ of all functions $f(x)$ defined for $a \leqslant x \leqslant b$ which are not necessarily continuous but **square-summable** (this means that the squares $f^2(x)$ are integrable in proper or improper sense; by the way, a function is said to be **summable** if it is absolutely integrable). The norm in the Hilbert

space is determined by the formula

$$\|f\| = \sqrt{\int_a^b [f(x)]^2\,dx} \tag{11}$$

and, according to the above condition, is always finite. This norm corresponds to the *mean square deviation* of functions.

It can easily be verified (we leave this to the reader) that all these norms satisfy Axioms 1-3.

To avoid confusion we write, when necessary, $\|f\|_{C[a, b]}$ and the like in order to indicate in what space the norm is considered. We once again note that the norm of every function entering into a normed functional space is finite. As an exercise, let the reader show that

$$\|x^2\|_{L_2[0, 1]} = \frac{1}{\sqrt{5}} \quad \text{and} \quad \left\|\frac{1}{x}\right\|_{L_2[0, 1]} = \infty$$

This means that the function $y = x^2$ belongs to the space $L_2[0, 1]$ or, in other words, is its **element, generalized vector**, while the function $y = \frac{1}{x}$ does not enter into $L_2[0, 1]$.

In a normed space we introduce the notion of a **distance** between any two elements f and g which we shall denote by ρ:

$$\rho(f, g) = \|f - g\|$$

The distance $\rho(f, g)$ between f and g defined in a functional space is nothing but the *deviation of the function $f(x)$ from the function $g(x)$ relative to the given norm.*

In courses of functional analysis it is proved that the spaces considered here possess a very important property referred to as **completeness.** Namely, any sequence $f_1, f_2, f_3, \ldots$ of elements for which $\|f_n - f_m\| \underset{\substack{n\to\infty \\ m\to\infty}}{\longrightarrow} 0$ has a limit in the corresponding space. Roughly speaking, in a complete space there is no sequence of elements of the space which converges, with respect to the norm of the space, to an element not belonging to the space. (We say that f_n converges to f and write $f_n \underset{n\to\infty}{\longrightarrow} f$ if $\|f_n - f\| \underset{n\to\infty}{\longrightarrow} 0$.) If, for instance, we take the collection of all continuous functions on a finite interval $a \leqslant x \leqslant b$ and introduce the norm according to formula (11), all the axioms of linear space and norm will be fulfilled but the space thus obtained will not be complete because if a sequence of continuous functions converges in the mean square its limit function may be discontinuous (e.g. see [107], Sec. XVII.9). Only after we take the **completion** of this space with respect to the norm, i.e. add to it all the mean square limits of the

sequences of continuous functions, we obtain a complete space (coinciding with the Hilbert space $L_2[a, b]$).

A complete normed linear space is called a **Banach space** after S. Banach (1892-1945), a famous Polish mathematician who was the first to carry out a thorough investigation of the properties of these spaces.

It should be noted that although every function belonging to $C_1[a, b]$ enters into $C[a, b]$ and every function from $C[a, b]$ is an element of $L_2[a, b]$ we cannot regard $C_1[a, b]$ as a subspace of $C[a, b]$ or $C[a, b]$ and $C_1[a, b]$ as subspaces of $L_2[a, b]$ because every normed space is considered together with its norm, and the norms of these spaces are different.

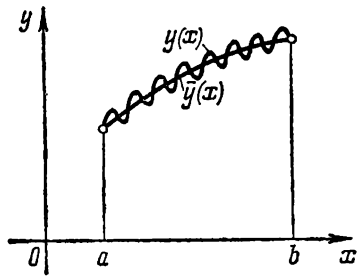


Fig. 79

When we deal with an extremum problem for a functional we must take into account the nature of the functional space involved. For instance, let a functional $I\{y\}$ have a local maximum for $\bar{y}(x)$. This means that for all functions $y(x)$ which do not identically coincide with $\bar{y}(x)$ and are sufficiently close to it we must have $I\{y\} < I\{\bar{y}\}$. But what is the meaning of the expression "functions sufficiently close to one another"? For example, is it legitimate to consider the function $y(x)$ shown in Fig. 79 as being close to $\bar{y}(x)$? The answer to the question obviously varies with the functional space in which $y(x)$ and $\bar{y}(x)$ are regarded as elements because there are different kinds of deviation of functions depending on the norm we deal with. Considering the graphs of the functions depicted in Fig. 79 we can say that $y(x)$ is close to $\bar{y}(x)$ if these functions are interpreted as elements of $C[a, b]$ but, at the same time, the deviation of $y(x)$ from $\bar{y}(x)$ is not small if we consider the functions to be elements of $C_1[a, b]$ (why?).

Accordingly, we consider different types of extrema of functionals. As a rule, in what follows we shall investigate functionals defined in C_1 or C. Then, if $I\{y\} < I\{\bar{y}\}$ for all the functions y which are sufficiently close to $\bar{y}$ in the sense of C, we say that the functional $I\{y\}$ has a **strong maximum** on the curve $y = \bar{y}(x)$. If $I\{y\} < I\{\bar{y}\}$ for all y whose deviation from $\bar{y}$ is small relative to the norm of C_1 we say that the functional $I\{y\}$ attains a **weak maximum** for the function $y = \bar{y}(x)$ (the notions of a **strong minimum** and **weak minimum** are introduced similarly).

Note that a strong extremum is always a weak one (why?) but a weak extremum is not necessarily a strong one. The latter is demonstrated by the example given below.

Suppose that a point moves in the plane in such a way that the absolute value v of its velocity $\mathbf{v}$ depends on the direction of motion according to the law shown graphically in Fig. 80 where α is the angle between the velocity vector $\mathbf{v}$ of the point and the positive direction of a fixed axis (l). Let it be necessary that the point should pass from a point A to a point B in the shortest time, the line segment AB being parallel to the axis (l) (see Fig. 81). To every possible path (S) connecting A with B

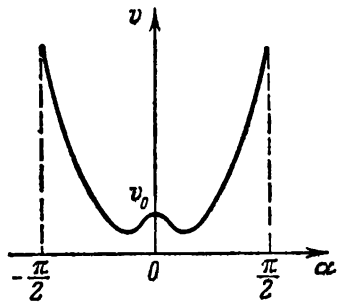


Fig. 80

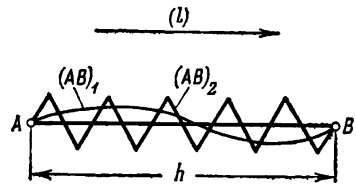


Fig. 81

there corresponds a time of motion T, and hence we have a minimum problem for the functional $T\{(S)\}$. It is readily seen that the rectilinear path (AB) gives $T\{(S)\}$ a weak minimum, the corresponding minimum time of motion being $\frac{h}{v_0}$ $(v_0 = v|_{\alpha=0})$. Indeed, for all the paths of the type of $(AB)_1$, which are close to (AB) both in their directions and relative to the distance between the points, the length of the path is greater while the speed is lower (see Figs. 80 and 81). But, at the same time, if the rate of growth of the function $v = v(\alpha)$ (whose graph is shown in Fig. 80) is sufficiently high for the values of α close to $\pm\frac{\pi}{2}$, the time of motion corresponding to the saw-tooth curves of the type of $(AB)_2$ (for which the direction of motion is changed, almost immediately, many times) can be less than $\frac{h}{v_0}$ (let the reader explain this fact!). A path of that type can lie as close as desired to (AB), and therefore in this example the weak minimum is not a strong one. If, in addition, the speed of motion decreases when the point goes away from the path (AB) (but, as before, the velocity of motion depends on α), we arrive at a curious situation: the narrower the "teeth" of the saw-tooth path, the shorter the time of motion. But of course these considera-

tions only serve as an approximate description of the situation because the direction of an actual motion cannot be changed during an arbitrarily small period of time, which has not been taken into account.

In conclusion we note that it is sometimes expedient to change the original functional space when it turns out that it yields no solution of the problem. For instance, if the original problem is considered in C_1 but proves to have no solution, one may try to pose the question as to whether it is possible to set this problem in the space C or in a wider space in such a way that the solution exists, investigate this solution (provided it exists), interpret it from the point of view of physics, etc. Such an approach may show some new features of the problem which were not known beforehand and facilitate its solution.

4. Variation of a Functional. The notion of the *variation* of a functional is very important for studying nonlinear functionals. Its role is similar to that of the differential of a nonlinear function. The differential of a nonlinear function is the principal part of its increment. This principal part is linear, and the replacement of the increment by the differential is equivalent to the linearization of the function for small variations of the argument (e.g. see [107], Secs. IV.8 and IX.11). Analogously, the *variation* of a nonlinear functional is equal to the principal part of its increment and is a linear functional. When we replace the increment of a functional, arising from the passage from one function on which the value of the functional depends to another function (close to the former), by the variation of the functional we thus linearize it.

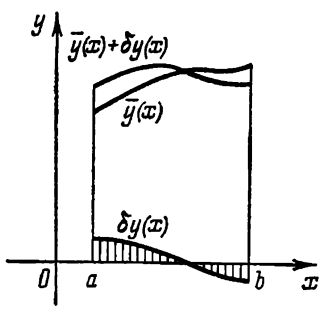


Fig. 82

As an example, let us take functional (8). First suppose that the function $y(x)$ on which the value of the functional depends coincides with a function $\bar{y}(x)$. Then let us pass to another function $\bar{y}(x)+\delta y(x)$ which is close to the former. Here $\delta y(x)$ (considered as being an indivisible symbol) is the **variation of the function** $y(x)$, that is an arbitrary function whose deviation from zero is small. This variation is added to the original function $\bar{y}(x)$ to obtain the new function $\bar{y}(x)+\delta y(x)$ (see Fig. 82). The deviation of $\delta y(x)$ from zero is understood in the sense of the norm (see Sec. 3) which has been chosen in the problem in question; for instance, in Fig. 82 the deviation is meant to be taken in the sense of the space C_1.

When $\bar{y}(x)$ is replaced by $\bar{y}(x)+\delta y(x)$ functional (8) gains the increment

$$\Delta I = I\{\bar{y}+\delta y\} - I\{\bar{y}\} = \int_0^2 [\bar{y}+\delta y]^2\,dx - \int_0^2 \bar{y}^2\,dx =$$

$$= 2\int_0^2 \bar{y}(x)\cdot\delta y(x)\,dx + \int_0^2 [\delta y(x)]^2\,dx \qquad (12)$$

Now let the function $\bar{y}(x)$ be fixed while the function $\delta y(x)$ be chosen in an arbitrary manner. Then the quantity ΔI is representable as a sum of two terms (formula (12)) each of which is a functional dependent on $\delta y = y - \bar{y}$. The first term

$$2\int_0^2 \bar{y}(x)\,\delta y(x)\,dx \qquad (13)$$

possesses the linearity property (see Sec. 2), while the second term

$$\int_0^2 [\delta y(x)]^2\,dx$$

is of higher order of smallness relative to small functions δy. Therefore, (13) is the principal part of increment (12) of functional (8) appearing when we pass from $\bar{y}$ to $\bar{y}+\delta y$ in the sense that it coincides with (12) to within the terms of higher order of smallness. Besides, this principal part is a linear functional. It is therefore expression (13) that is called the *(first) variation δI of functional* (8). Hence, denoting the variation by δI we can write

$$\delta I = \delta\int_0^2 y^2\,dx = 2\int_0^2 y(x)\,\delta y(x)\,dx$$

where the original notation $\bar{y}(x)$ is replaced by $y(x)$ because this function was taken quite arbitrarily.

The above example is typical. In the general case we are given a functional $I\{y\}$ in which we pass from a function $y(x)$ to a new function $y(x)+\delta y(x)$ and thus obtain the increment ΔI of the functional. As a rule, the increment

$$\Delta I = I\{y+\delta y\} - I\{y\}$$

is representable in the form

$$\Delta I = I_1\{y,\ \delta y\} + R_1\{y,\ \delta y\} \qquad (14)$$

where the first summand is a linear functional dependent on $\delta y(x)$ (for a fixed function $y(x)$) while the second summand is of higher

order of smallness relative to $\delta y(x)$. Then the first term on the right-hand side of (14) is called the **(first) variation of the functional** I and we write

$$\delta I = I_1\{y,\ \delta y\}$$

If the terms of higher order of smallness are negligibly small we simply say that the variation of a functional is (approximately) equal to its infinitesimal increment gained when the function on which the functional depends receives an infinitesimal variation. The replacement of the increment of a functional by its variation is nothing but the **linearization** of the functional. All these considerations are completely analogous to those connected with the introduction of the notion of the differential of a function (e.g. see [107], Secs. IV.7, 8).

In concrete problems the variation of a functional can be found by means of Taylor's formula (e.g. see [107], Secs. IV.15 and XII.6). For instance, consider the functional

$$I\{y\} = \int_a^b F(x,\ y)\,dx \tag{15}$$

where, under the sign of integration, the quantity y is regarded as being dependent on x, i.e. as a function $y = y(x)$. Then we have

$$\Delta I = \int_a^b [F(x,\ y+\delta y)\,dx - F(x,\ y)]\,dx \tag{16}$$

Now, substituting the expression

$$F(x,\ y+\delta y) = F(x,\ y) + F'_y(x,\ y)\,\delta y + \textit{infinitesimals of higher order of smallness}$$

into (16) and discarding the terms of higher order of smallness we finally obtain

$$\delta I = \delta \int_a^b F(x,\ y)\,dx = \int_a^b F'_y(x,\ y)\,\delta y\,dx \tag{17}$$

For the functional

$$I\{y\} = \int_a^b F(x,\ y,\ y')\,dx \tag{18}$$

we similarly obtain (check it up!) the formula

$$\delta I = \int_a^b [F'_y(x,\ y,\ y')\,\delta y + F'_{y'}(x,\ y,\ y')\,\delta y']\,dx \tag{19}$$

where $\delta y'$ can be interpreted both as $\delta(y')$ and $(\delta y)'$ because the derivative of a difference of two functions is equal to the difference of their derivatives.

5. More Precise Definition. To give a more precise description of the terms entering into resolution (14) we must specify the functional space (see Sec. 3) in which (or in part of which) the functional in question is considered. For instance, it appears natural to regard functional (15) as being defined in $C[a, b]$, functional (18) in $C_1[a, b]$ and so on. In some cases the choice of the space is made on the basis of some physical considerations or in accord with certain mathematical traditions, etc. Now suppose that the functional space (and thus the type of the norm $\|f\|$) has been already chosen. Then the second summand in (14) must satisfy, for a fixed function $y(x)$, the condition

$$|R_1\{y;\ \delta y\}| = o(\|\delta y\|) \quad (\text{for } \|\delta y\| \to 0) \tag{20}$$

which means that it is of higher order of smallness relative to δy. Practically, we usually introduce the stronger condition

$$|R_1\{y;\ \delta y\}| = O(\|\delta y\|^2) \quad (\text{for } \|\delta y\| \to 0)$$

that is require that R_1 be of the second or higher order of smallness relative to δy.

For a chosen function $y(x)$, the first summand entering into (14), which is linear in δy, must satisfy the inequality

$$|I_1\{y,\ \delta y\}| \leqslant K\|\delta y\| \tag{21}$$

It should be noted that we can require that relation (21) be fulfilled for all the admissible variations δy or that it must be satisfied by all δy with a sufficiently small norm because, by the first linearity property (see (9)), when δy is multiplied by a constant C both sides of (21) are multiplied by $|C|$.

A linear functional satisfying condition of type (21) is said to be **bounded**, and the least value of the constant K for which relation (21) holds is referred to as the **norm** of the linear functional. This norm is a positive number for all bounded functionals except the zero functional (whose all values are zero) for which it assumes the zero value. For instance, it can readily be shown that the first term on the right-hand side of (12), regarded as a functional dependent on $\delta y \in C[0, 2]$ (for a fixed function $\bar{y}(x)$), is a bounded linear functional whose norm is equal to

$$2\int_0^2 |\bar{y}(x)|\,dx$$

(let the reader prove this result).

Consider another example. Let the functional

$$I_1\{y\} = 3y'(0) + \int_0^1 xy(x)\,dx \tag{22}$$

which is linear with respect to $y(x)$ be considered in the space $C_1[0, 1]$. Then it is bounded because we have

$$|I_1\{y\}| \leqslant 3 \max_{0 \leqslant x \leqslant 1} |y'(x)| + \max_{0 \leqslant x \leqslant 1} |y(x)| \int_0^1 x\,dx \leqslant 3\|y\|_{C_1[0,\ 1]}$$

(see (10)). But if the same expression (22) is regarded as being defined in the space $C[0, 1]$ it is no longer a bounded linear functional. Indeed, taking the sequence of functions $y_n(x) = \frac{1}{n} \sin nx$ $(n = 1, 2, 3, \ldots)$ we can write

$$I_1\{y_n\} = 3 + \int_0^1 \frac{x}{n} \sin nx\,dx \xrightarrow[n \to \infty]{} 3, \quad \|y_n\|_{C[0,\ 1]} \xrightarrow[n \to \infty]{} 0$$

But this condition cannot hold for a bounded functional (why?). Moreover, for the function $y = \sqrt{x}$ (which belongs to $C[0, 1]$ but does not belong to $C_1[0, 1]$!) functional (22) has no finite value.

This example reveals a typical situation. In courses of functional analysis it is proved that *if a linear functional is defined in a Banach space* (R) (which means that at each point of (R) it assumes a finite value) *it is bounded.*

In conclusion let us discuss the possibility of linearizing a functional. As is known, in mathematical analysis we sometimes deal with continuous functions $f(x)$ which are not differentiable for some values of x. Examples of such functions are $y = \sqrt[3]{x}$, $y = \sqrt[3]{x^2}$ and $y = |x|$. These functions are continuous but not differentiable at the point $x = 0$ (remember the shape of their graphs), which means that they cannot be linearized at this point. Similarly, a functional $I\{y\}$ defined in a functional space does not necessarily admit of linearization for all the points of the functional space, i.e. for all functions $y(x)$ belonging to the space, even if it is **continuous** *) (that is such that, for the chosen norm, its value gains small increments when the variations of $y(x)$ are sufficiently small).

For example, let us consider the functional

$$I\{y\} = \int_0^1 \sqrt[3]{y}\,dx \tag{23}$$

in the space $C[0, 1]$. It is continuous but cannot be linearized in the vicinity of the "point" (that is function) $y(x) \equiv 0$, $0 \leqslant x \leqslant 1$

*) It is proved in functional analysis that a linear functional defined in a normed functional space is continuous if and only if it is bounded.— *Tr.*

(why?). Moreover, functional (23) cannot be linearized in the vicinity of every function which vanishes at all the points of an arbitrary subinterval of the interval [0, 1]. It can also be shown that it cannot be linearized in the vicinity of every function possessing a zero of order not less than $\frac{3}{2}$ while the presence of zeros of order less than $\frac{3}{2}$ does not violate the possibility of linearization.

If we take a more general functional of type (15) (with continuous function $F(x, y)$) regarded in the space $C[a, b]$ formula (17) indicates that the possibility of linearization can be violated at points and on curves (provided there are such) where $F'_y(x, y)$ has discontinuities. If $y=\varphi(x)$ is the equation of such a singular curve, the functional cannot be linearized for every function coinciding with $\varphi(x)$ at all points of an interval. If $F'_y(x, y)$ approaches infinity for $y-\varphi(x)\to 0$ and is of the order of $[y-\varphi(x)]^{-p}$, the linearization is impossible for every function $y(x)$ for which the expression $y(x)-\varphi(x)$ has at least one zero of order equal to or higher than $\frac{1}{p}$.

Here we also mention another approach to the notion of variation. Suppose that we have a functional, for instance, of type (15), in which a function of a particular form $y=\varphi(x; \lambda)$ (involving a scalar parameter λ) is substituted for $y(x)$. It is meant that the parameter λ can be varied arbitrarily, which results in a change in the functional relationship $y=y(x)$. Then, for this class of functions, the functional I becomes a function of λ, i.e. $I=I(\lambda)$. Performing the linearization we obtain

$$\delta y=\partial_\lambda \varphi \doteq \frac{\partial \varphi}{\partial \lambda} d\lambda, \quad \delta I=d_\lambda I=\frac{dI}{d\lambda} d\lambda \tag{24}$$

This result makes it possible to derive formula (17) directly (do it!). But in this approach we must take into account that if we regard the variations δy as quite arbitrary, we must consider the one-parameter family of functions $\varphi(x, \lambda)$ to be arbitrary as well. A similar approach is possible when the functional relationship $y(x)$ involves more than one parameter.

6. Necessary Condition for Extremum. The necessary condition for extremum of a functional we are going to derive in this section is completely analogous to the well-known necessary condition for extremum of a function of one or several independent variables. Suppose that a function $\bar{y}(x)$ gives a local maximum or minimum to a functional $I\{y\}$ in a chosen functional space (R), this functional possessing the variation $\delta I\{\bar{y}, \delta y\}$, i.e. admitting of linearization in the vicinity of $y=\bar{y}(x)$. Besides, let us suppose that

this extremum is not one-sided in the sense that the functional $I\{y\}$ is defined for all the functions $\bar{y}$ which are sufficiently close, relative to the norm of the space under consideration, to the function y (e.g. cf. [107], Secs. IV.19 and XII.11). In what folfows, unless otherwise stated, we shall suppose that this condition is always fulfilled.

Then, *for the extremum to exist, it is necessary that the condition*

$$\delta I\{\bar{y},\ \delta y\}=0 \tag{25}$$

be fulfilled for any $\delta y \in (R)$.

In fact, let, for definiteness, the functional I have a minimum for $y=\bar{y}(x)$. Suppose that $\delta I\{\bar{y},\ \delta y\}>0$ for a function δy. Substituting $k\delta y$ for δy into (14) (where k is a scalar) we obtain

$$\Delta I = I\{\bar{y}+k\delta y\}-I\{\bar{y}\}=I_1\{\bar{y},\ k\delta y\}+R_1\{\bar{y},\ k\delta y\}=$$
$$=k\left(I_1\{\bar{y},\ \delta y\}+\frac{1}{k}R_1\{\bar{y},\ k\delta y\}\right)$$

But, for small values of $|k|$, the left member must be positive while the sign of the right member coincides with that of k (prove this assertion by taking into account the inequality which must hold for the term R_1), and thus it can be both positive and negative. We thus arrive at a contradiction which proves the necessity of condition (25).

Condition (25) can also be derived from the second formula (24). Indeed, if $\bar{y}(x)=\varphi(x,\ \bar{\lambda})$, then the function $I(\lambda)$ has an extremum for $\lambda=\bar{\lambda}$ whence $\left.\frac{dI}{d\lambda}\right|_{\lambda=\bar{\lambda}}=0$, which implies (25).

As was mentioned in Sec. 1, in concrete problems we often consider an extremum of a functional on condition that its values are compared not for all functions of a given functional space but for a narrower class of functions satisfying certain nonhomogeneous linear subsidiary relations of the form

$$y(a)=y_a,\ y(b)=y_b \tag{26}$$

where the values y_a and y_b are given. Then condition (25) must be fulfilled for every variation δy satisfying the corresponding homogeneous relations which, in the case of conditions (26), are of the form

$$\delta y(a)=0,\ \delta y(b)=0 \tag{27}$$

Indeed, for these δy the functions of the form $y+k\delta y$ also satisfy conditions (26) (why?), and therefore the proof of condition (25) given above applies to this case as well.

As is known (e.g. see [107], Sec. VII.19), a nonhomogeneous linear relation determines a *hyperplane* in a (linear) vector space. In the case of a functional space we have the same situation, but the hyperplane we are speaking about consists of elements of an infinite-dimensional space. A variational problem considered for a functional $I\{y\}$, whose values assumed on a chosen class of functions are compared, is thus regarded on a curvilinear manifold (S) of a functional space (R). The linearization in the vicinity of the "point of extremum" $\bar{y} \in (S)$ shows that condition (25) must be fulfilled for every function δy belonging to the "tangent hyperplane" drawn to (S) at the point $\bar{y}$.

7. Euler's Equation. The necessary condition of form (25) makes it possible to determine the sought-for solution $\bar{y}(x)$ in many problems. But sometimes formula (25) is not convenient because it involves not only $\bar{y}(x)$ but also an arbitrary function δy. Therefore, condition (25) is usually transformed to an equivalent relation which only contains the unknown function $\bar{y}(x)$. The way in which this transformation is performed depends on the particular class of functionals under consideration, and the main object of § 1 is to describe these classes and the corresponding necessary conditions for extremum. As a rule, the necessary condition derived in this way reduces to an equation (known as **Euler's equation**) and some *subsidiary (boundary) conditions*. Euler's equation is usually a differential equation (although it sometimes turns into a finite equation). As for the subsidiary conditions, some of them can be given in the original formulation of the problem while the others are derived from condition (25).

Let us begin with some simple special cases.

By virtue of formula (17), for an extremum of the functional

$$I\{y\} = \int_a^b F(x, y)\, dx \tag{28}$$

the necessary condition (25) takes the form

$$\int_a^b F'_y(x, \bar{y}(x))\, \delta y\, dx = 0 \tag{29}$$

If, for definiteness, the problem is considered for all continuous functions $y(x)$, i.e. in the space $C[a, b]$, condition (29) must hold for any continuous function $\delta y(x)$. Thus, it follows that

$$F'_y(x, \bar{y}(x)) \equiv 0 \tag{30}$$

Indeed, denoting the left member of (30) as $\varphi(x)$ and putting $\delta y = \varphi(x)$ (which is legitimate since δy is an arbitrary continuous

function) we obtain, by (29), the relation

$$\int_a^b [\varphi(x)]^2 dx = 0 \tag{31}$$

whence $\varphi(x) \equiv 0$ (why?), which yields (30).

Thus, *Euler's equation for functional* (28) *has the form*

$$F'_y(x, y) = 0 \tag{32}$$

Condition (32) can also be obtained if we approximately replace integral (28) by the corresponding integral sum

$$\sum_{k=1}^{n} F(\xi_k, \bar{y}_k) \Delta x_k \qquad (\bar{y}_k = \bar{y}(\xi_k)) \tag{33}$$

The function $\bar{y}(x)$ being arbitrary, we can vary the value $\bar{y}_k$ $(k = 1, 2, \ldots, n)$ considering the other values of the function entering into (33) as being fixed. Then, if sum (33) attains an extremal value, its derivative with respect to $\bar{y}_k$ must be equal to zero, which results in (32).

In some problems we encounter functionals of form (18):

$$I\{y\} = \int_a^b F(x, y, y')\, dx$$

We shall compare the values of this functional for all continuously differentiable functions $y(x)$ (i.e. for the elements y of the space $C_1[a, b]$) satisfying the given boundary conditions (26):

$$y(a) = y_a, \quad y(b) = y_b$$

On the basis of formulas (19) and (25) we conclude that the relation

$$\int_a^b [F'_y(x, \bar{y}(x), \bar{y}'(x))\, \delta y + F'_{y'}(x, \bar{y}(x), \bar{y}'(x))\, \delta y']\, dx = 0 \tag{34}$$

must be fulfilled for every function $\delta y \in C_1[a, b]$ satisfying conditions (27).

To derive the equation for the function $\bar{y}(x)$ let us integrate by parts the second term in (34):

$$\int_a^b F'_{y'}(x, \bar{y}(x), \bar{y}'(x))\, \delta y'\, dx =$$

$$= [F'_{y'}(x, \bar{y}(x), \bar{y}'(x))\, \delta y]_{x=a}^{x=b} - \int_a^b [F'_{y'}(x, \bar{y}(x), \bar{y}'(x))]'\, \delta y\, dx =$$

$$= -\int_a^b [F'_{y'}(x, \bar{y}(x), \bar{y}'(x))]'\, \delta y\, dx \tag{35}$$

(here, in accordance with condition (27), the substitution $[F'_{y'}(x, \bar{y}, \bar{y}')\delta y]\Big|_{x=a}^{x=b}$ vanishes). Substituting (35) into (34) we obtain

$$\int_a^b \{F'_y(x, \bar{y}(x), \bar{y}'(x)) - [F'_{y'}(x, \bar{y}(x), \bar{y}'(x))]'\}\, \delta y\, dx = 0$$

The function δy being arbitrary, it follows immediately that the expression in the curly brackets is identically equal to zero for $a \leqslant x \leqslant b$. Thus we arrive at the differential equation

$$F'_y - \frac{d}{dx} F'_{y'} = 0 \tag{36}$$

where the symbol $\frac{d}{dx}$ designates the operation of taking the total derivative (i.e. when performing the differentiation we take into account the dependence of y and y' on x). To prove the result stated let us designate temporarily the expression in the curly brackets by $\varphi(x)$. If $\varphi(a) = \varphi(b) = 0$ we can simply put $\delta y = \varphi(x)$ and apply the argument used in the deduction of formula (30), which yields (36). In the general case we can put $\delta y = \varphi(x)$ on an interval of the form $a + \varepsilon \leqslant x \leqslant b - \varepsilon$ where $\varepsilon > 0$ is very small and define the function δy on the intervals $a \leqslant x \leqslant a + \varepsilon$ and $b - x \leqslant x \leqslant b$ in such a way that it is continuous on the entire interval $a \leqslant x \leqslant b$ and assumes zero values at its end points. Then, making ε tend to zero we obviously arrive at relation (31) which in this case implies (36).

Thus, *for functional* (18) *considered under conditions* (26), *Euler's equation takes the form* (36).

Writing in full the expression for the total derivative by applying the rule for differentiating a composite function and replacing $\bar{y}$ by y we obtain

$$F'_y(x, y, y') - F''_{xy'}(x, y, y') - F''_{yy'}(x, y, y')\, y' - F''_{y'y'}(x, y, y')\, y'' = 0 \tag{37}$$

(check up the calculations!). We see that in the case of functional (18) Euler's equation is an ordinary second-order differential equation (if $F''_{y'y'} \equiv 0$, that is when F is a linear function of y', the term $-F''_{y'y'}y''$ vanishes and (37) becomes of lower order; this degenerate case is of no interest and we shall not consider it here). Every solution of Euler's equation (37) is called an **extremal** of functional (18). Every extremal gives the functional a **stationary value** in the sense that if $y(x)$ is an extremal defined on an interval $a_1 \leqslant x \leqslant b_1$, $a \leqslant a_1 < b_1 \leqslant b$ (in particular, the interval $[a_1, b_1]$ may coincide with $[a, b]$), and we give $y(x)$ an arbitrary

variation on that interval without changing the values $y(a_1)$ and $y(b_1)$, then

$$\delta \int_{a_1}^{b_1} F(x, y, y')\,dx = 0$$

This assertion is proved by means of the same transformations as in the foregoing paragraph (let the reader check it up). The term "extremal" is also applied to the graphs of the solutions of equation (37), that is to the integral curves of (37).

The collection of all extremals of functional (18) is the general solution of the ordinary second-order differential equation (36) and, consequently, it is a two-parameter family of functions. Thus, boundary conditions (26) are sufficient for the determination of the sought-for particular solution. But it should be noted that Euler's equation can rarely be integrated by quadratures, and therefore in problem solving practice we have to resort to some numerical methods of solving boundary-value problems (which we shall not discuss here).

We see that extremals can be equivalently defined in two different ways, namely, as integral curves of Euler's equation and as curves giving the functional stationary values. This implies, in particular, the following **invariance property** of Euler's equation. Suppose that we pass, in functional (18), from the variables x, y to new variables $X = X(x, y)$, $Y = Y(x, y)$. Then we obtain a new functional possessing its own Euler's equation. But, according to the above, the lines of stationarity of the former functional go into the lines of stationarity of the latter functional under this change of variables, and therefore Euler's equation for the original functional is transformed into Euler's equation for the new functional.

8. Examples. There are important particular cases when Euler's equation (36) has a *first integral* (e.g. on this notion see [107], Sec. XV.13) and thus can be reduced to an ordinary differential equation of the first order. For instance, this is the case when the function F in functional (18) does not involve (explicitly) the variable x. Virtually, if $F'_x \equiv 0$, the second term in equation (37) vanishes identically, and, after both sides are multiplied by y', the equation takes the form (check it up!)

$$[F(y, y') - y'F'_{y'}(y, y')]'_x = 0$$

Integrating we obtain the first integral

$$F(y, y') - y'F'_{y'}(y, y') = C_1 \tag{38}$$

where C_1 is an arbitrary constant.

We suggest that the reader investigate the case when the function F is independent of y, that is has the form $F=F(x, y')$, and find the first integral.

Let us consider the second example in Sec. 1 in which it is required to minimize functional (4) under boundary conditions (5). This is just an example of a functional of type (18) with $F=F(y, y')$. Here, making use of first integral (38), we obtain (check up the calculations!)

$$\frac{1}{\sqrt{h-y}\sqrt{1+y'^2}}=C_1 \tag{39}$$

This equation can be transformed to the form $y'=f(y)$ after which the variables can be separated. But it is more convenient to make the substitution

$$h-y=\frac{1-\cos t}{2C_1^2} \qquad \left(dy=-\frac{\sin t}{2C_1^2}\,dt\right) \tag{40}$$

Then, after some simple calculations that we leave to the reader we obtain, on the basis of (39), the relations

$$dx=\pm\frac{1}{2C_1^2}(1-\cos t)\,dt, \quad x=\frac{1}{2C_1^2}(t-\sin t)+C_2 \tag{41}$$

(here we may omit the signs $\pm$ because t can always be replaced by $-t$). Equations (40), (41) determine a *cycloid* (e.g. see [107],

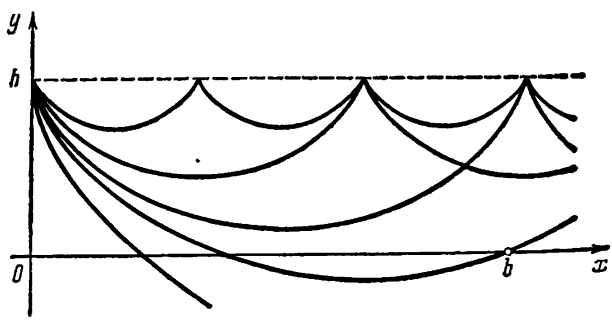


Fig. 83

formula (II.12)) which is the trajectory of a fixed point on the circumference of a circle of radius $R=\frac{1}{2C_1^2}$, when the circle rolles upon the straight line $y=h$. The sought-for curve must pass through the point $(0, h)$. Therefore, there must be a *cuspidal point* of the cycloid at that point, and hence we can put $C_2=0$. The family of these cycloids is depicted in Fig. 83. The value of the radius R which is arbitrary and has not yet been chosen should be such

that the second boundary condition is satisfied, that is the cycloid must pass through the point $(b, 0)$. Substituting these coordinates into (40) and (41) and eliminating t we readily obtain a transcendental equation for R. If the arc of the cycloid corresponding to a given value of R (and to the range $0 \leqslant t \leqslant 2\pi$ of the parameter t) is constructed, the sought-for value of R can be found by means of a similitude transformation (how?). Here there is an interesting special case when $b > \pi h$. Then a part of the sought-for curve lies under the terminal point, the result not being obvious. The explanation of this fact is that when the whole trajectory is long the moving point must descend a sufficient distance in order to gain a high speed so that it could travel along the trajectory in the shortest time and arrive at the terminal point lying at the required level.

Taking the third example of Sec. 1 and applying integral (38) in a similar manner we obtain the equation

$$\frac{y}{\sqrt{1+y'^2}} = C_1$$

(check it up!). The integration of the equation (which we leave to the reader) results in

$$y = C_1 \cosh \frac{x - C_2}{C_1} \tag{42}$$

The symmetry considerations indicate that $C_2 = 0$. The curve thus determined (for $C_2 = 0$) is obtained from the graph of the function $y = \cosh x$ by means of the similitude transformation with the ratio of similitude C_1. It is called a **catenary**, and the corresponding surface of revolution is termed a **catenoid.**

9. Functionals Involving Derivatives of Higher Order. Let us take a functional of the form

$$I\{y\} = \int_a^b F(x, y, y', y'')\, dx \tag{43}$$

with boundary conditions

$$y(a) = y_a, \quad y'(a) = y'_a, \quad y(b) = y_b, \quad y'(b) = y'_b \tag{44}$$

where the values y_a, y'_a, y_b and y'_b are given. The variation of functional (43) is expressed by the formula

$$\delta I = \int_a^b (F'_y \delta y + F'_{y'} \delta y' + F'_{y''} \delta y'')\, dx \tag{45}$$

Using the necessary condition for extremum (formula (25)) and reasoning as in Sec. 7 (here the last term in (45) should be integrated by parts twice) we can obtain the corresponding Euler equation.

Performing these calculations, we conclude that *Euler's equation for functional* (43) *is of the form*

$$F'_y - \frac{d}{dx} F'_{y'} + \frac{d^2}{dx^2} F'_{y''} = 0$$

This is an ordinary differential equation of the fourth order whose general solution (the collection of all the extremals) involves four arbitrary parameters. Hence, four boundary conditions (44) are sufficient (at least in principle) for determining the sought-for particular solution $y(x)$ which realizes the extremum.

Functionals containing derivatives up to order $k > 2$ inclusive are treated quite analogously; Euler's equation for a functional of this type is an ordinary differential equation of the $2k$th order.

10. Functionals of Several Functions. In the calculus of variations we also consider functionals which depend on several functions of one independent variable. In particular, this is the case when the sought-for line is a space curve (why?). On the other hand, problems of this kind may arise irrespective of their geometrical interpretation.

Consider a functional of the form

$$I\{y_1, y_2\} = \int_a^b F(x, y_1, y'_1, y_2, y'_2)\, dx \tag{46}$$

dependent on two functions $y_1(x)$ and $y_2(x)$ defined for $a \leqslant x \leqslant b$. It is supposed that these functions are completely independent of each other, and therefore we can fix one of them and vary, in an arbitrary manner, the other. Fixing $y_2(x)$ and giving $y_1(x)$ an arbitrary variation we arrive at the corresponding variation of the functional:

$$\delta_{y_1} I\{y_1, y_2, \delta y_1\} = \int_a^b \left(F'_{y_1} \delta y_1 + F'_{y'_1} \delta y'_1\right) dx \tag{47}$$

Expression (47) is an analogue of a partial differential of a function (e.g. see [107], Sec. IX.10). Interchanging the roles of y_1 and y_2 we similarly obtain

$$\delta_{y_2} I\{y_1, y_2, \delta y_2\} = \int_a^b \left(F'_{y_2} \delta y_2 + F'_{y'_2} \delta y'_2\right) dx \tag{48}$$

If we have a variational problem for functional (46) with the simplest boundary conditions

$$y_1(a) = y_{1a}, \quad y_1(b) = y_{1b}, \quad y_2(a) = y_{2a}, \quad y_2(b) = y_{2b} \tag{49}$$

where the values y_{1a}, y_{1b}, y_{2a} and y_{2b} are given, both variations (47) and (48) must be equal to zero.

Hence, reasoning as in Sec. 7, we conclude that *in the case of functional* (46) *the necessary condition for extremum is expressed by the system of Euler's equations*

$$F'_{y_1}-\frac{d}{dx}F'_{y_1'}=0, \quad F'_{y_2}-\frac{d}{dx}F'_{y_2'}=0 \tag{50}$$

The general solution of this system of two ordinary differential equations of the second order contains four arbitrary constants which are found from the four boundary conditions.

Functionals involving an arbitrary number of functions and also those containing derivatives of higher order of several functions are investigated in like manner.

If we introduce the vector function

$$\mathbf{y}(x)=\begin{pmatrix} y_1(x) \\ y_2(x) \end{pmatrix}$$

of the scalar argument x (e.g. on vector functions see [107], Secs. VII.23 and XI.1), functional (46) can be written as the expression

$$I\{\mathbf{y}\}=\int_a^b F(x, \mathbf{y}, \mathbf{y}')\,dx \tag{51}$$

dependent on that vector function. It appears natural to introduce the total variation of the functional equal to the sum of partial variations (47) and (48); after some simple transformations we write it in the form

$$\delta I\{\mathbf{y}, \delta\mathbf{y}\}=\int_a^b (F'_{\mathbf{y}}\cdot\delta\mathbf{y}+F'_{\mathbf{y}'}\cdot\delta\mathbf{y}')\,dx$$

where $F'_{\mathbf{y}}$ designates the vector

$$F'_{\mathbf{y}}=\begin{pmatrix} F'_{y_1} \\ F'_{y_2} \end{pmatrix} \tag{52}$$

(which is the gradient of the function F with respect to the variables y_1 and y_2; e.g. see [107], Sec. XII. 1), and the expressions $F'_{\mathbf{y}}\cdot\delta\mathbf{y}$ and $F'_{\mathbf{y}'}\cdot\delta\mathbf{y}'$ are scalar products. Boundary conditions (49) and the system of Euler's equations (50) can also be written in the vector form

$$\mathbf{y}(a)=\mathbf{y}_a, \quad \mathbf{y}(b)=\mathbf{y}_b, \quad F'_{\mathbf{y}}-\frac{d}{dx}F'_{\mathbf{y}'}=\mathbf{0}$$

We see that all these relations are completely analogous to the formulas obtained in the case of scalar functions. The vector form of Euler's equations is very convenient because it holds for any number of functions on which the functional is dependent. Besides, there are some problems in which an unknown vector function

is involved in the original formulation of the problem while the scalar functions (its projections on the coordinate axes) only appear after a Cartesian basis has been introduced in the space of vectors y. In these problems the vector form is particularly preferable because it is invariant with respect to the choice of the basis.

Functionals of several functions are directly related to **variational problems in parametric form.** Suppose that we are given a functional of form (18) or a more general functional

$$I((L)) = \int_{(L)} F(x, y, y')\, dx \tag{53}$$

where (L) is an arbitrary oriented curve. It is often convenient to represent this curve parametrically in the form $x = x(t)$, $y = y(t)$, which yields

$$I((L)) = \int_{t_{\text{initial}}}^{t_{\text{terminal}}} F\left(x, y, \frac{\dot{y}}{\dot{x}}\right)\dot{x}\, dt = \int_{t_{\text{initial}}}^{t_{\text{terminal}}} F_1(x, y, \dot{x}, \dot{y})\, dt \tag{54}$$

Thus we arrive at a new functional of type (46). Characteristic features of the function F_1 in (54) are that it does not involve explicitly the new independent variable t and is a homogeneous function of degree one with respect to $\dot{x}$ and $\dot{y}$ (e.g. on the definition of a homogeneous function of a given degree see [107], Sec. IX.12). Conversely, it can readily be shown that if the function F_1 involved into a functional of form (54) possesses these properties, expression (54). can be reduced to form (53). Euler's equations entering into the system corresponding to (54) turn out to be dependent (check it up by differentiating the Euler identity $F'_{1\dot{x}}\dot{x} + F'_{1\dot{y}}\dot{y} = F$). Therefore, the general solution of the system contains an arbitrary function depending on the way in which the curve (L) has been parametrized. Choosing a parametrization of (L) we thus eliminate this function; for example, we can introduce the additional equality $\dot{x} = 1$ which means that it is the coordinate x that has been taken as the parameter t; similarly, the equality $\dot{x}^2 + \dot{y}^2 = 1$ means that the parameter chosen is the arc length reckoned along (L) and the like.

Functionals dependent on curves in spaces of higher dimension are investigated in a similar way.

11. Functionals of Functions of Several Variables. A functional may also involve a function of several independent variables. For definiteness, let us consider the class of functions $z(x, y)$ of two independent variables x and y defined in a domain (G) of

the xy-plane, i.e. $(x, y) \in (G)$. Then we can take the functional

$$I\{z\} = \int_{(G)} F\left(x, y, z, \frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right) dG \tag{55}$$

which is an analogue of (18). The variation of functional (55) is written in the form

$$\delta I\{z, \delta z\} = \int_{(G)} \left(F_z' \delta z + F_p' \delta \frac{\partial z}{\partial x} + F_q' \delta \frac{\partial z}{\partial y}\right) dG \tag{56}$$

where p and q designate the partial derivatives $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$, and the expressions $\delta \frac{\partial z}{\partial x}$ and $\delta \frac{\partial z}{\partial y}$ coincide, respectively, with $\frac{\partial (\delta z)}{\partial x}$ and $\frac{\partial (\delta z)}{\partial y}$. The symbols F_p' and F_q' denote the partial derivatives of F with respect to $p = \frac{\partial z}{\partial x}$ and $q = \frac{\partial z}{\partial y}$.

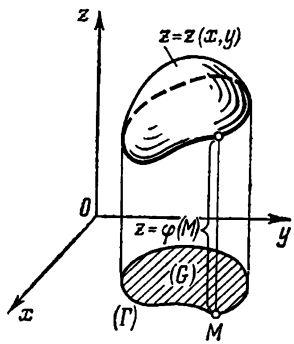


Fig. 84

The boundary condition (which is a generalization of (26)) is taken in the form

$$z|_{M \in (\Gamma)} = \varphi(M) \qquad (M = M(x, y)) \tag{57}$$

where (Γ) is the contour bounding the domain (G), $M(x, y)$ is the variable point on the curve (Γ) and $\varphi(M)$ is a given function defined on (Γ). Hence, condition (57) means that the values of the function $z(x, y)$ assumed on the contour (Γ) are given. The geometric significance of this condition is that we consider all the possible surfaces, described by equations of the form $z = z(x, y)$, pulled over the fixed space curve whose equation is (57) (see Fig. 84 where this space curve is shown in heavy line).

To derive the necessary condition for extremum of functional (55) under boundary condition (57) we make use of general condition (25) and, like in Sec. 7, perform integration by parts according to the scheme

$$\int_{(G)} F_p' \frac{\partial (\delta z)}{\partial x} dG = \int dy \int F_p' \frac{\partial (\delta z)}{\partial x} dx - \int dy \left[(F_p' \delta z)\Big|_{(\Gamma)} - \int \frac{\partial F_p'}{\partial x} \delta z \, dx \right] \tag{58}$$

where the symbol $(F_p' \delta z)|_{(\Gamma)}$ designates the double substitution of boundary values of x into $F_p' \delta z$. By analogy with Sec. 7, we conclude, taking into account condition (57), that $(\delta z)|_{(\Gamma)} = 0$, and therefore the result of the double substitution is zero. Hence, the right-hand side of (58) is equal to $-\int_{(G)} \frac{\partial F_p'}{\partial x} \delta z \, dG$.

Transforming the last term in (56) in the same manner we see that *Euler's equation for functional* (55) *has the form*

$$F'_z - \frac{\partial}{\partial x} F'_p - \frac{\partial}{\partial y} F'_q = 0 \tag{59}$$

This equation was first obtained by M.V. Ostrogradsky in 1834 and therefore is also referred to as **Ostrogradsky's equation.** In what follows we shall call it the **Euler-Ostrogradsky equation.** It should be noted that the symbols $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ involved in (59) designate the operations of taking the *total partial derivatives* with respect to x and y; for instance, when applying the operation $\frac{\partial}{\partial x}$ we differentiate the function F'_p for a fixed value of y but take into account that the expressions z, $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$ entering into F'_p are dependent on x because they are functions of x (and y). Consequently, writing in full equation (59) we obtain (check it up!)

$$F''_{pp} \frac{\partial^2 z}{\partial x^2} + 2F''_{pq} \frac{\partial^2 z}{\partial x \partial y} + F''_{qq} \frac{\partial^2 z}{\partial y^2} + F''_{zp} \frac{\partial z}{\partial x} + F''_{zq} \frac{\partial z}{\partial y} + F''_{xp} + F''_{yq} - F'_z = 0$$

This is a partial differential equation of the second order, and thus the variational problem in question reduces to solving this equation with boundary condition (57). In this course we shall not consider the general theory of partial differential equations.

The case when the unknown function depends on more than two independent variables, say on $x_1, x_2, \ldots, x_n$, is treated in the same way (in problems of this kind the notation $\frac{\partial z}{\partial x_k} = p_k$, $k = 1, 2, \ldots, n$, is usually applied).

A more general variational problem with a functional of the form

$$I\{z\} = \int_{(G)} F\left(x, y, z, \frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}, \frac{\partial^2 z}{\partial x^2}, \frac{\partial^2 z}{\partial x \partial y}, \frac{\partial^2 z}{\partial y^2}\right) dG$$

and the boundary conditions

$$z\Big|_{M \in (\Gamma)} = \varphi(M), \quad \frac{\partial z}{\partial n}\Big|_{M \in (\Gamma)} = \psi(M)$$

is investigated by analogy with (55). Here (Γ) is again the boundary of (G), and n designates the direction of the inner normal to (Γ), $\varphi(M)$ and $\psi(M)$ being given functions on (Γ). Let the reader prove that the corresponding Euler-Ostrogradsky equation is of the form

$$F'_z - \frac{\partial}{\partial x} F'_p - \frac{\partial}{\partial y} F'_q + \frac{\partial^2}{\partial x^2} F'_r + \frac{\partial^2}{\partial x \partial y} F'_s + \frac{\partial^2}{\partial y^2} F'_t = 0$$

where the symbols r, s and t denote, respectively, the second-order par-

tial derivatives $\frac{\partial^2 z}{\partial x^2}$, $\frac{\partial^2 z}{\partial x \partial y}$ and $\frac{\partial^2 z}{\partial y^2}$. When establishing this result one must take into account that the equalities $(\delta z)_{(\Gamma)} = 0$ and $\left(\frac{\partial}{\partial n} \delta z\right)_{(\Gamma)} = 0$ imply the relations $\left(\frac{\partial}{\partial x} \delta z\right)_{(\Gamma)} = \left(\frac{\partial}{\partial y} \delta z\right)_{(\Gamma)} = 0$.

If, like in Sec. 10, the expression z in (55) is a k-dimensional vector function, the corresponding Euler-Ostrogradsky equation retains form (59) in which z, p and q should be interpreted as vectors, and the differentiation with respect to a vector is understood in the sense of (52). This vector equation is equivalent to the corresponding system of k (scalar) partial differential equations of the second order.

12. Isoperimetric Problems. Let us come back to functional (18) and suppose that the unknown function is not quite arbitrary but must satisfy, besides boundary condition (26), an additional condition of the form

$$J\{y\} = \int_a^b G(x, y, y')\, dx = g \tag{60}$$

where $G(x, y, y')$ is a given function and the value g of the integral is also given. Then the variation δy must also satisfy the additional condition

$$\delta J\{y, \delta y\} = \int_a^b (G'_y \delta y + G'_{y'} \delta y')\, dx = 0$$

which is obtained from the linearization of (60).

Suppose that a function $\bar{y}(x)$ realizes a (conditional) extremum of functional (18) or its (conditional) stationary value under boundary condition (26) and *integral constraint* (60). Then we can reason as follows. Let us break up the interval $a \leqslant x \leqslant b$ into a large number n of equal parts $h = \frac{b-a}{n}$ and put $x_k = a + kh$, $\bar{y}_k = y(x_k)$. Replacing the integrals by their integral sums and the derivatives by the divided differences we see that the problem we are interested in can be approximately reduced to determining a conditional extremum for the function

$$I_n = \sum_{k=0}^{n-1} F\left(x_k, \bar{y}_k, \frac{\bar{y}_{k+1} - \bar{y}_k}{h}\right) h \tag{61}$$

of the independent variables $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{n-1}$ with given values $\bar{y}_0 = y_a$, $\bar{y}_n = y_b$ under the subsidiary condition

$$\sum_{k=0}^{n-1} G\left(x_k, \bar{y}_k, \frac{\bar{y}_{k+1} - \bar{y}_k}{h}\right) h = g \tag{62}$$

(This argument is not, of course, a rigorous proof; it is only aimed at elucidating the situation in principle.)

We now deal with a function of a finite number of independent variables $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{n-1}$ and therefore can apply the theory of conditional extremum for functions of several variables (e. g. see [107], Sec. XII.10). According to this theory we must consider the unconditional extremum for the function

$$\sum_{k=0}^{n-1} F\left(x_k, \bar{y}_k, \frac{\bar{y}_{k+1}-\bar{y}_k}{h}\right) h - \lambda \sum_{k=0}^{n-1} G\left(x_k, \bar{y}_k, \frac{\bar{y}_{k+1}-\bar{y}_k}{h}\right) h \qquad (63)$$

where λ is **Lagrange's multiplier.** Returning now from the integral sums to the integrals we see that the function $\bar{y}(x)$ giving functional (18) a conditional extremum satisfies Euler's equation written for the function

$$F^*(x, y, y', \lambda) = F(x, y, y') - \lambda G(x, y, y') \qquad (64)$$

(This final result can also be derived by considering the original problem (18), (60) without resorting to integral sums.) The general solution of Euler's equation corresponding to function (64) contains two arbitrary constants and the unknown parameter λ. These unknown quantities are found from two boundary conditions (26) and relation (60) (the constraint). It can be shown (e.g. see considerations given in [107], Sec. XII.10) that the Lagrange's multiplier λ must take on the value $\frac{dI(\bar{y})}{dg}$.

If we have several conditions of type (60), that is

$$\int_a^b G_i(x, y, y')\, dx = g_i \qquad (i = 1, 2, \ldots, p) \qquad (65)$$

then, instead of (64), we must take the functional with

$$F^* = F - \sum_{i=1}^{p} \lambda_i G_i$$

and similarly consider its unconditional extremum.

Now let us return to the first example of Sec. 1 which reduces to maximizing functional (1) with boundary conditions (3) and integral constraint (2) expressing the constancy of the perimeter L. By the way, by analogy with this example, the variational problems with integral subsidiary conditions of form (65) or other similar conditions (referred to as **isoperimetric conditions**) are termed **isoperimetric problems.** According to the above, we must solve Euler's equation corresponding to the function

$$F^*(x, y, y', \lambda) = y - \lambda\sqrt{1+y'^2} \qquad (66)$$

(This argument is not, of course, a rigorous proof; it is only aimed at elucidating the situation in principle.)

We now deal with a function of a finite number of independent variables $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{n-1}$ and therefore can apply the theory of conditional extremum for functions of several variables (e. g. see [107], Sec. XII.10). According to this theory we must consider the unconditional extremum for the function

$$\sum_{k=0}^{n-1} F\left(x_k, \bar{y}_k, \frac{\bar{y}_{k+1}-\bar{y}_k}{h}\right) h - \lambda \sum_{k=0}^{n-1} G\left(x_k, \bar{y}_k, \frac{\bar{y}_{k+1}-\bar{y}_k}{h}\right) h \qquad (63)$$

where λ is **Lagrange's multiplier.** Returning now from the integral sums to the integrals we see that the function $\bar{y}(x)$ giving functional (18) a conditional extremum satisfies Euler's equation written for the function

$$F^*(x, y, y', \lambda) = F(x, y, y') - \lambda G(x, y, y') \qquad (64)$$

(This final result can also be derived by considering the original problem (18), (60) without resorting to integral sums.) The general solution of Euler's equation corresponding to function (64) contains two arbitrary constants and the unknown parameter λ. These unknown quantities are found from two boundary conditions (26) and relation (60) (the constraint). It can be shown (e.g. see considerations given in [107], Sec. XII.10) that the Lagrange's multiplier λ must take on the value $\frac{dI(\bar{y})}{dg}$.

If we have several conditions of type (60), that is

$$\int_a^b G_i(x, y, y')\, dx = g_i \qquad (i = 1, 2, \ldots, p) \qquad (65)$$

then, instead of (64), we must take the functional with

$$F^* = F - \sum_{i=1}^{p} \lambda_i G_i$$

and similarly consider its unconditional extremum.

Now let us return to the first example of Sec. 1 which reduces to maximizing functional (1) with boundary conditions (3) and integral constraint (2) expressing the constancy of the perimeter L. By the way, by analogy with this example, the variational problems with integral subsidiary conditions of form (65) or other similar conditions (referred to as **isoperimetric conditions**) are termed **isoperimetric problems.** According to the above, we must solve Euler's equation corresponding to the function

$$F^*(x, y, y', \lambda) = y - \lambda\sqrt{1+y'^2} \qquad (66)$$

This function does not involve x, and therefore we can take advantage of first integral (38), which yields

$$y-\frac{\lambda}{\sqrt{1+y'^2}}=C_1$$

Solving this equation with respect to y' we obtain

$$\frac{dy}{dx}=\pm\frac{\sqrt{\lambda^2-(y-C_1)^2}}{y-C_1}$$

Now, separating the variables and integrating we arrive at the equation

$$(x-C_2)^2+(y-C_1)^2=\lambda^2$$

describing the family of circles of radii λ with centres at the points (C_1, C_2). The constants C_1, C_2 and λ are found from boundary conditions (3) and constraint (2). In other words, among the circular arcs shown in Fig. 85 we must choose the one having the given length L.

When considering the arcs of the type of $aABCb$ along which y is not a single-valued function of x we have an ambiguous situation because in the above argument we have essentially used the assumption that y is a one-valued function. But, as in many similar cases, here we can resort to the following consideration. Suppose that a curve of the type of $aABCb$ (which has not yet been determined) gives the functional the sought-for extremum. Take an arbitrary small arc of the curve and choose new coordinates ξ and η such that the equation of this arc is written in the form $\eta=\eta(\xi)$ where $\eta(\xi)$ is a single-valued function (see Fig. 85). Now imagine that we fix the points A and B and vary the arc AB while its length is retained constant. This cannot lead to an increase of the enclosed area of the lot (why?). Thus we come to the following conclusion (which, as can be shown, expresses a general property of variational problems): *if a curve gives an extremum to a functional in a variational problem its every arc is also a solution of an analogous variational problem.* (This assertion obviously remains true when we deal with a stationary value of a functional.) The function $\eta(\xi)$ being one-valued, the curve AB is an arc of a circle. But if every small part of a line (L) is a circular arc the whole line (L) is also an arc of a circle, which is what we set out to prove.

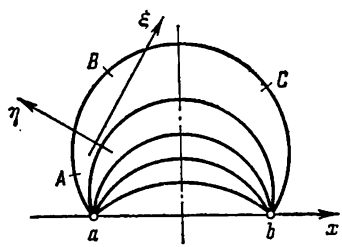


Fig 85.

We can also rely upon the invariance of Euler's equation mentioned at the end of Sec. 7. Indeed, it is always possible to in-

troduce curvilinear coordinates u, v, in which the entire line under consideration is represented by a single-valued function $v=v(u)$. Then this line must be an integral curve of Euler's equation in the new coordinates, and hence, by the invariance property, in the old coordinates as well.

Now consider the problem of finding a line of minimum length bounding a given area. This is the **reciprocal problem** of the problem studied above. It turns out that a circular arc is the solution of the reciprocal problem as well. This can be established directly by solving the corresponding Euler equation but we shall resort to the following general consideration. Let the given area be equal to S, and (L) be the corresponding line of minimum length L distinct from the circular arc. Then, according to the original problem, the curve (L) can be replaced by a line (L') of length $L'=L$ bounding an area $S'>S$ and lying close to (L). Now we can deform (L') in such a way that it goes into a curve (L'') while its length and the area it bounds are decreased (for instance, we can perform a uniform contraction toward the x-axis). This deformation can be performed so that the area S'' bounded by (L'') again becomes equal to S: $S''=S$. But then we have $L''<L'$, i. e. $L''<L$, which contradicts the definition of the line (L).

The above situation demonstrates the general **reciprocity principle** for variational problems. In the general case it becomes more evident if we speak about stationary values of functional (18) with constraint (60) and of functional (60) with constraint of type (18). In fact, the former problem reduces to the solution of Euler's equation corresponding to the function $F-\lambda_1 G$ and the latter to the solution of Euler's equation corresponding to $G-\lambda_2 F$. But we have

$$G-\lambda_2 F=-\lambda_2\left(F-\frac{1}{\lambda_2}G\right)=-\lambda_2(F-\lambda_1 G)\qquad\left(\lambda_1=\frac{1}{\lambda_2}\right)$$

and consequently, after the cancellation is made, we arrive at the same equation as in the former problem. (In these problems the values $\lambda_2=\frac{dJ(\bar{y})}{dI(\bar{y})}=0$ and $\lambda_1=\frac{dI(\bar{y})}{dJ(\bar{y})}=0$ of the parameters λ_2 and λ_1 are singular because they lead to the degeneration of the problems.) It should be noted that the reciprocity principle also remains valid for extremal problems containing several unknown functions.

Variational problems with isoperimetric constraints are widely encountered in modern applied mathematics. For instance, we often deal with a situation in which certain resources must be allocated so as to achieve the maximum profit. If the possible ways in which the decision-maker can design the allocation are described by a function or a set of functions we usually arrive

at a variational problem in which the conditions determining the amounts of the resources at hand are expressed in the form of integral constraints.

13. Conditional Extremum with Finite and Differential Constraints. For definiteness, let us take an extremum problem for the functional

$$I\{y_1, y_2, y_3\} = \int_a^b F(x, y_1, y_1', y_2, y_2', y_3, y_3')\,dx \qquad (67)$$

dependent on three unknown functions $y_1(x)$, $y_2(x)$ and $y_3(x)$. Suppose that a constraint of the form

$$H(x, y_1, y_2, y_3) = h(x) \qquad (68)$$

(where the functions H and h are given) is imposed on these functions. In this case we speak about a **finite** or **geometric** constraint (in mechanics also called a **holonomic** constraint). Then the variations of the functions must satisfy the obvious relation

$$H_{y_1}'\delta y_1 + H_{y_2}'\delta y_2 + H_{y_3}'\delta y_3 = 0$$

Suppose we are looking for a (conditional) extremal value or a (conditional) stationary value of the functional (67) under finite constraint (68) and boundary conditions

$$y_k(a) = y_{ka}, \quad y_k(b) = y_{kb}, \qquad (k = 1, 2, 3)$$

coherent with the equation of the constraint, that is

$$H(a, y_{1a}, y_{2a}, y_{3a}) = h(a), \quad H(b, y_{1b}, y_{2b}, y_{3b}) = h(b)$$

(This is the **compatibility condition** of boundary conditions with constraint.)

If equation (68) makes it possible to express one of the sought-for functions in terms of the others, for instance, to find the representation $y_3 = \varphi(x, y_1, y_2)$, we can substitute it into (67) and thus pass to an unconditional variational problem with the two unknown functions $y_1(x)$ and $y_2(x)$ (see Sec. 10). But in the general case such a reduction is impossible, and then we can apply the argument used in Sec. 12, which in this case, instead of (62), results in the relations

$$H(x_k, \bar{y}_{1k}, \bar{y}_{2k}, \bar{y}_{3k}) = h(x_k) \qquad (k = 1, 2, \ldots, n-1) \qquad (69)$$

According to the rule of determining a conditional extremum we must take, for each $k = 1, 2, \ldots, n-1$, a Lagrange multiplier λ_k, i.e. consider the unconditional extremum for the function

$$\sum_{k=0}^{n-1} F\left(x_k, \bar{y}_k, \frac{\bar{y}_{1,k+1} - \bar{y}_{1k}}{h}, \ldots\right) h - \sum_{k=1}^{n-1} \lambda_k H(x_k, \bar{y}_{1k}, \bar{y}_{2k}, \bar{y}_{3k})$$

After the necessary conditions for the extremum have been written we return, from the integral sums, to the integrals and thus conclude that to determine the unknown functions $\overline{y}_1(x)$, $\overline{y}_2(x)$, $\overline{y}_3(x)$ giving functional (67) the conditional extremum we must solve the system of Euler's equations corresponding to the function

$$F(x, y_1, y_1', y_2, y_2', y_3, y_3') - r(x) H(x, y_1, y_2, y_3) \qquad (70)$$

where the function $r(x)$ is an analogue of Lagrange's multiplier (which was constant in the foregoing problems). The function $r(x)$ is not known in advance but is considered to be fixed, that is it does not receive a variation when we consider the increment of the functional. The corresponding Euler equations together with constraint equation (68) form a system of four equations in four unknown functions $y_1(x)$, $y_2(x)$, $y_3(x)$ and $r(x)$. After it has been integrated, the arbitrary constants appearing in its general solution are determined from the boundary conditions.

Similarly, if there are two independent finite constraints of the form

$$H_1(x, y_1, y_2, y_3) = h_1(x), \quad H_2(x, y_1, y_2, y_3) = h_2(x)$$

imposed on the functions y_1, y_2, y_3 we must take, instead of (70), the function

$$F^* = F - r_1(x) H_1 - r_2(x) H_2$$

and consider the corresponding Euler equations. (It should be noted that in contrast to integral constraints the number of finite constraints cannot exceed two if we have three unknown functions.)

If, in addition, the problem in question involves a number of integral constraints (65) we must add to the function F^* the sum $-\sum_i \lambda_i G_i$.

Now we consider a constraint of the form

$$H(x, y_1, y_1', y_2, y_2', y_3, y_3') = h(x)$$

where H and h are given functions. It contains the derivatives y_1', y_2', y_3' and therefore is called a **differential (nonholonomic)** constraint. In this case we must also add to the function F the term $-r(x) H$. The distinction from the foregoing problem lies in the fact that now the system of Euler's equations involves not only $r(x)$ but also $r'(x)$ (why?). A variational problem with finite and differential constraints is usually referred to as **Lagrange's problem.**

As an example, let us consider the problem of S.A. Chaplygin in which it is required to determine the shape of the closed path along which an airplane must fly in a given time T on condition that the wind velocity v is constant (in its magnitude and direc-

tion) and that the area bounded by the path is maximum. Take a Cartesian coordinate system x, y, in which the y-axis goes along the direction of the wind, and denote by V the relative velocity of the airplane. If $x=x(t)$ and $y=y(t)$ are the equations of motion (which simultaneously specify a parametric representation of the sought-for trajectory), we must have

$$\dot{x}^2+(\dot{y}-v)^2=V^2 \tag{71}$$

where the dot designates the differentiation with respect to time t. For the area bounded by the path we have the well-known formula

$$S=\frac{1}{2}\int_0^T (x\dot{y}-y\dot{x})\,dt \tag{72}$$

(e.g. see [107], formula (XIV.31)), and besides we can write the obvious conditions $x(0)=x(T)=x_0$, $y(0)=y(T)=y_0$. We thus arrive at the problem of maximizing integral (72) with differential constraint (71). According to the above, we must form the system of Euler's equations for the function

$$F^*=\frac{1}{2}(x\dot{y}-y\dot{x})-r(t)\,[\dot{x}^2+(\dot{y}-v)^2]$$

This system is written (check it up!) in the form

$$\frac{1}{2}\dot{y}-\frac{d}{dt}\left[-\frac{1}{2}y-r(t)\cdot 2\dot{x}\right]=0,\quad -\frac{1}{2}\dot{x}-\frac{d}{dt}\left[\frac{1}{2}x-r(t)\cdot 2(\dot{y}-v)\right]=0$$

Integrating with respect to t we obtain

$$y+2r(t)\dot{x}=C_1,\quad -x+2r(t)(\dot{y}-v)=C_2 \tag{73}$$

Translating the coordinate axes we can always place them so that $C_1=C_2=0$ (why?). For simplicity's sake, let us assume that this transformation of the axes has already been made. Eliminating $r(t)$ from (73) we receive

$$\frac{-y}{\dot{x}}=\frac{x}{\dot{y}-v}\quad\text{whence}\quad dt=\frac{x\,dx+y\,dy}{vy}$$

Substituting this expression into (71) we get the differential equation

$$v^2(x^2+y^2)\,dx^2=V^2(x\,dx+y\,dy)^2 \tag{74}$$

which involves solely the quantities x and y. This equation is homogeneous in x and y (e.g. see [107], Sec. XV.4, Case 2). But it can be solved in a simpler manner than by means of the general methods if we pass to polar coordinates ρ, φ. Indeed, substituting $x=\rho\cos\varphi$ and $y=\rho\sin\varphi$ into (74) and extracting the square root we derive (check it up!) the relation

$$\frac{d\rho}{\rho}=\frac{\pm v\sin\varphi}{V\pm v\cos\varphi}\,d\varphi=\frac{\pm\varepsilon\sin\varphi}{1\pm\varepsilon\cos\varphi}\,d\varphi\qquad\left(\varepsilon=\frac{v}{V}\right)$$

Integrating and taking the antilogarithms we find the equation of the sought-for family of curves:

$$\rho = \frac{C}{1 \pm \varepsilon \cos \varphi}$$

Here C is an arbitrary constant, and the equation obtained is the polar equation of the family of ellipses (because, according to the sense of the problem, we have $\varepsilon < 1$). These ellipses have the same eccentricity $\varepsilon = \frac{v}{V}$, their major axes being perpendicular to the direction of the wind.

For functionals dependent on functions of several arguments there sometimes appear (besides finite, differential and integral constraints) subsidiary conditions which are, in a certain sense, combinations of these types of constraints. For instance, let the function $z(x, y)$ in functional (55) be subject to a constraint of the form

$$\int G\left(x, y, z, \frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right) dy = g(x) \tag{75}$$

where the integration with respect to y (the integral is one-fold here) is performed within the limits specified by the shape of the domain (G). Applying arguments similar to those at the beginning of this section we conclude that in this case Euler's equation must be formed for the function $F^* = F - r(x) G$; this yields an equation in two unknown functions $z(x, y)$ and $r(x)$ which must be considered together with the integro-differential equation (75) of the constraint.

14. Problems Reducible to Lagrange's Problem. There are various classes of variational problems with conditional extrema reducible to Lagrange's problem. Here we shall enumerate some of these problems which can serve as patterns for transforming problems of other types.

1. Take the extremum problem for functional (18) with integral constraint (60) which was investigated in Sec. 12. This problem is reduced to Lagrange's problem if we introduce the new unknown function

$$z(x) = \int_a^x G(s, y(s), y'(s)) ds$$

This results in the extremum problem for a functional of form (18) under the differential constraint $z' = G(x, y, y')$ and boundary conditions

$$y(a) = y_a, \quad y(b) = y_b, \quad z(a) = 0, \quad z(b) = g$$

2. Let it be required to find two functions $y_1(x)$ and $y_2(x)$, connected by an equation

$$f(x,\ y_1,\ y_1',\ y_2,\ y_2') = 0 \tag{76}$$

and satisfying boundary conditions

$$y_1(a) = y_{1a},\ y_2(a) = y_{2a},\ y_2(b) = y_{2b} \tag{77}$$

such that the quantity $y_1(b)$ assumes an extremal value in comparison with the other functions satisfying the same relations. (This is a particular case of the general problem referred to as **Mayer's problem**). Introducing the new unknown function

$$y_3 = y_1' \tag{78}$$

we reduce this problem to Lagrange's problem of extremizing the functional

$$\int_a^b y_3\,dx$$

with differential constraints (76) and (78) and boundary conditions (77). (At first glance it seems that the number of boundary conditions in the problem thus obtained is insufficient for determining the unknown quantities, but in Secs. 15 and 16 we shall elaborate some methods of solving variational problems in which the number of the boundary conditions involved in their formulations is less than in the problems studied before.)

Mayer's problem with an arbitrary number n of unknown functions connected by a system of m differential equations (it is supposed that $m < n$; why is this condition introduced?) is investigated in like manner.

Let the reader consider an analogous problem of determining an extremal value of the quantity $y(b)$ for the solution of a given differential equation $y' = f(x,\ y,\ \varphi(x))$ containing an arbitrary function $\varphi(x)$ (which, for example, may be a coefficient entering into the equation and the like) satisfying a subsidiary relation of the form

$$\int_a^b G(x,\ y,\ \varphi)\,dx = g$$

where the initial value $y(a) = y_a$ is given.

3. In the **problem of Bolza** it is required to extremize a **functional**

$$\int_a^b F(x,\ y,\ y',\ \lambda)\,dx + \varphi(y(a),\ y(b),\ \lambda) \tag{79}$$

containing a parameter λ (in the general case such a problem may involve any number of parameters, unknown functions and subsidiary conditions of the type of finite, differential and integral constraints). Let the reader show that this problem is reduced to Lagrange's problem for extremum of the functional

$$\int_a^b [F(x, y_1, y_1', y_2) + y_3]\, dx$$

under the differential constraints

$$y_2' = 0, \quad y_3' = 0$$

and the additional boundary condition.

$$(b-a)\, y_3(a) - \varphi(y_1(a), y_1(b), y_2(a)) = 0 \tag{80}$$

In connection with (80) we note that a boundary condition may connect the values of the unknown functions assumed at both end points of the interval on which they are considered.

In [122] the reader can find the solution of two-dimensional and three-dimensional problems in which a functional analogous to (79) is of the form of a sum of integrals taken over the domain in question and over its boundary.

15. Extremals with Moving Boundaries in the Plane. Let us come back to functional (18). Up till now we only considered boundary conditions of form (26). In other words, we regarded the end points of the curve $y = y(x)$ as being fixed. But there are variational problems with boundary conditions of a more general type which we are going to discuss here.

Suppose that the values of functional (18) are compared for the functions $y(x) \in C_1[a, b]$ satisfying the boundary conditions

$$f_1(y(a), y'(a)) = 0, \quad f_2(y(b), y'(b)) = 0 \tag{81}$$

which are a generalization of (26). This means that we do not set the ordinate at each end point but introduce a relation between this ordinate and the slope of the tangent to the curve. Let $y = \bar{y}(x)$ be a function which gives the sought-for extremum to the functional. Then, if we take a narrower class of functions satisfying the boundary conditions

$$y(a) = \bar{y}(a), \quad y'(a) = \bar{y}'(a), \quad y(b) = \bar{y}(b), \quad y'(b) = \bar{y}'(b)$$

and compare the values of the functional only for these functions, the extremum will again be attained for $y = \bar{y}(x)$ because conditions (81) hold for these functions. Consequently, we can give $\bar{y}(x)$ an arbitrary variation defined on the interval $a \leqslant x \leqslant b$ and reason as in Sec. 6. This enables us to draw the conclusion that the function $\bar{y}(x)$ must satisfy Euler's equation (36). But in this

case the two arbitrary constants involved in the general solution of Euler's equation must be determined with the aid of two boundary conditions of form (81).

It is clear that the above argument is of a general character and therefore a function which gives an extremum to a functional is a solution of the corresponding Euler equation for any boundary conditions.

Now let us consider a variational problem with the so-called *natural boundary conditions*. Suppose that the boundary conditions are dropped and that the values of functional (18) are compared for all the functions $y(x) \in C_1[a, b]$. Then, first of all, reasoning as above, we see that the function $\bar{y}(x)$ for which the extremum is reached satisfies Euler's equation. Taking an arbitrary variation δy we obtain, integrating by parts, the expression

$$\delta I\{\bar{y}, \delta y\} = \int_a^b (F'_y \delta y + F'_{y'} \delta y') dx =$$

$$= F'_{y'} \delta y \Big|_{x=a}^{b} + \int_a^b \left(F'_y - \frac{d}{dx} F'_{y'}\right) \delta y \, dx$$

The last term on the right-hand side being equal to zero, we thus obtain the relation

$$\delta I\{\bar{y}, \delta y\} = F'_{y'}(b, \bar{y}(b), \bar{y}'(b)) \delta y(b) - F'_{y'}(a, \bar{y}(a), \bar{y}'(a)) \delta y(a)$$

Condition (25) is always necessary and the quantities $\delta y(a)$, $\delta y(b)$ are quite arbitrary; therefore we conclude that the sought-for function $\bar{y}(x)$ satisfies the boundary conditions

$$F'_{y'}(a, y(a), y'(a)) = 0, \quad F'_{y'}(b, y(b), y'(b)) = 0 \qquad (82)$$

which are called **natural**. Of course, if the boundary condition is removed only on one end point the corresponding natural boundary condition of type (82) is only imposed at that end point.

Let the reader consider a more general functional of the form

$$\int_a^b F(x, y, y') dx + \varphi_1(y)\Big|_{y=a} + \varphi_2(y)\Big|_{y=b} \qquad (83)$$

and show that in this case the natural boundary conditions are of the form

$$(F'_{y'} - \varphi'_1)\big|_{y=a} = 0, \quad (F'_{y'} + \varphi'_2)\big|_{y=b} = 0 \qquad (84)$$

16. Transversality Conditions. Let us consider a more general variational problem for functional (18) in which the values a and b are not fixed beforehand, and the terminal points $(a, y(a))$ and $(b, y(b))$ of the sought-for curve are constrained only to lie on

given lines whose equations are, respectively,

$$f_{\mathrm{I}}(x, y)=0 \quad \text{and} \quad f_{\mathrm{II}}(x, y)=0 \tag{85}$$

(see Fig. 86). It should be stressed that, generally speaking, the functions $y(x)$, for which the values of the functional are compared, no longer form a linear space (why?). But the approach to the notion of the variation of a functional indicated at the end of Sec. 5 remains valid although the formulas for the variation are appropriately changed because it is necessary to take into account the possibility of variation of the limits of integration. Performing the linearization we obtain the expression

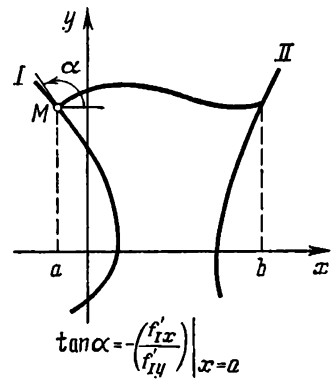


Fig. 86

$$\delta\int_a^b F(x, y, y')\,dx=\int_a^b (F'_y\,\delta y+F'_{y'}\,\delta y')\,dx+$$
$$+F(x, y, y')|_{x=b}\,\delta b-F(x, y, y')|_{x=a}\,\delta a$$

Reasoning as at the beginning of Sec. 15 we conclude that the function $y(x)$ for which the functional attains the sought-for extremum satisfies Euler's equation. Therefore, applying necessary condition (25) we derive from it, with the aid of integration by parts, the relation

$$[F'_{y'}\,\delta y]|_{x=b}-[F'_{y'}\,\delta y]|_{x=a}+F|_{x=b}\,\delta b-F|_{x=a}\,\delta a=0 \tag{86}$$

On the other hand, equations (85) imply the equalities

$$\left.\begin{aligned} f'_{Ix}|_{x=a}\,\delta a+f'_{Iy}|_{x=a}\,\delta[y(a)]=0 \\ f'_{IIx}|_{x=b}\,\delta b+f'_{IIy}|_{x=b}\,\delta[y(b)]=0 \end{aligned}\right\} \tag{87}$$

But one must note that the quantities $\delta y|_{x=b}$ and $\delta[y(b)]$ do not coincide in the general case! Considering Fig. 87 we see that after the infinitesimals of higher order of smallness have been discarded we arrive at the relation

$$\delta[y(b)]=\delta y|_{x=b}+y'|_{x=b}\,\delta b \tag{88}$$

and at a similar relation for the quantities $\delta[y(a)]$ and $\delta y|_{x=a}$. Substituting these expressions into (87) we find $\delta y|_{x=a}$ and $\delta y|_{x=b}$ after which, taking into account that the quantities δa and δb can assume arbitrary values, we obtain the relations (check up the calculations!)

$$\left.\begin{aligned} [F'_{y'}(f'_{Ix}+y'f'_{Iy})-Ff'_{Iy}]|_{x=a}=0 \\ [F'_{y'}(f'_{IIx}+y'f'_{IIy})-Ff'_{IIy}]|_{x=b}=0 \end{aligned}\right\} \tag{89}$$

Thus, in this case the role of the natural boundary conditions is played by the equalities (89). If the direction of line I (or II), i.e. of its tangent, is known at its point M (Fig. 86), this specifies the value of the ratio $f'_x : f'_y$, and therefore conditions (89) determine the direction of the extremal (of its tangent line) at that point. An extremal of functional (18) which satisfies a condition of type (89) relative to a given curve (curve I or II in our case) is termed a **transversal**. Accordingly, conditions (89) are called **transversality conditions.**

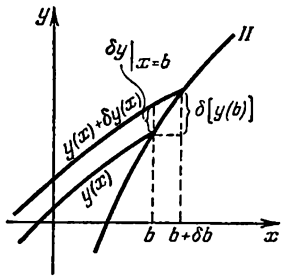


Fig. 87

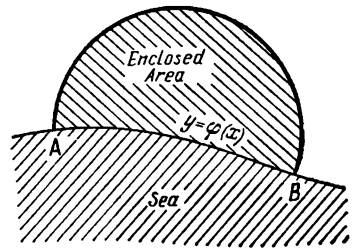


Fig. 88

By the way, note that conditions (82) are a particular case of conditions (89) (why?).

Thus, at every point $M(x, y)$ of the xy-plane, to each direction (specified by a slope k), there corresponds a set of directions with slopes l satisfying, at that point, the transversality condition

$$(k-l)F'_{y'}(x, y, l)+F(x, y, l)=0 \qquad (90)$$

from which the values of l can be found.

For example, let us consider the following variant of Dido's problem (Sec. 1). Suppose that the end points A and B of the thread are situated on the coastline whose equation is $y=\varphi(x)$ where $\varphi(x)$ is an arbitrary (given) function, the positions of A and B not being fixed beforehand (see Fig. 88). Then integral (1) expressing the enclosed area is replaced by the integral

$$S=\int_a^b [y-\varphi(x)]\,dx$$

Thus, we have $F=y-\varphi(x)$, and expression (66) is replaced by $F^*=y-\varphi(x)-\lambda\sqrt{1+y'^2}$. Obviously, this change does not affect the form of Euler's equation, and transversality condition (90) is written as

$$(k-l)\frac{-\lambda l}{\sqrt{1+l^2}}+y-\varphi(x)-\lambda\sqrt{1+l^2}=0$$

This condition must be fulfilled at the points of the curve $y=\varphi(x)$, and hence we obtain

$$-(k-l)\lambda l-\lambda(1+l^2)=0, \quad \text{that is} \quad kl+1=0, \quad l=-\frac{1}{k}$$

The latter relation is nothing but the well-known perpendicularity condition for two directions (e.g. see [107], Sec. II.9, Problem 5). Consequently, if the positions of the end points of the thread are not fixed beforehand, the tangents to the thread are normal to the coastline at the end points.

Transversality condition (90) has been established for unconditional extremum problems, while the above problem involves an integral constraint. But it can be easily shown that the necessary conditions for extremum in problems involving extremals with moving end points (and other similar problems) are extended to problems with integral, finite and differential constraints without any changes and must be applied to the corresponding functionals modified by means of Lagrange's multipliers.

Finally, when deriving the transversality conditions, we have incidentally obtained the general expression

$$\delta\int_a^b F(x,\ y,\ y')\,dx=\int_a^b\left(F'_y-\frac{d}{dx}F'_{y'}\right)\delta y\,dx+F'_{y'}|_{x=b}\,\delta[y(b)]+$$
$$+(F-y'F'_{y'})|_{x=b}\,\delta b-F'_{y'}|_{x=a}\,\delta[y(a)]-(F-y'F'_{y'})|_{x=a}\,\delta a \quad (91)$$

for the variation of functional (18) which applies to arbitrary variations of the function $y(x)$ and arbitrary limits of integration. In particular, formula (91) can be used for extremals with free boundaries encountered in problems without subsidiary conditions on one or both end points that is when one or both end points can move in an arbitrary manner. Applying formula (91) to an end point of that kind we arrive at the equalities

$$F=0, \qquad F'_{y'}=0$$

These two equalities (like the equation of the constraint and the transversality condition in the above problem) together with the expressions of y and y' obtained from the general solution of Euler's equation result in one relation (corresponding to this end point) connecting two arbitrary constants entering into the general solution.

17. Extremals with Moving Boundaries in Space. Let us consider functional (46) which can be interpreted as being dependent on a line in the xy_1y_2-space. In Sec. 10 we supposed that the end points of the line were fixed because the values of x, y_1, and y_2 were given at each end point. The case when we have three inde-

pendent equations connecting the values of x, y_1, y_2, y_1', y_2' at each end point is treated analogously.

But there are more general problems in which we have only one or two conditions at an end point (usually there is at least one such condition although we can also consider problems in which a spatial curve has free boundaries). Here we shall limit ourselves to the case of finite constraints connecting the coordinates (but not their derivatives). Both end points are considered in the same way, and therefore, for the sake of simplicity, we shall consider the moving end point $x=b$. Note that the considerations given in Sec. 15 indicate that the functions $y_1(x)$ and $y_2(x)$ giving the sought-for extremum to the functional must satisfy the corresponding system of Euler's equations.

To begin with, suppose that the end point $x=b$ has two degrees of freedom, that is there is only one constraint of the form

$$f(x, y_1, y_2)=0 \tag{92}$$

connecting the coordinates of the end point. By analogy with relations (86) and (87), we conclude that the relation

$$f_x'|_b\,\delta b+f_{y_1}'|_b\,(\delta y_1|_b+y_1'|_b\,\delta b)+f_{y_2}'|_b\,(\delta y_2|_b+y_2'|_b\,\delta b)=0$$

is fulfilled, and it must imply the equality

$$\left[F_{y_1'}'\delta y_1\right]_{x=b}+\left[F_{y_2'}'\delta y_2\right]_{x=b}+F|_{x=b}\,\delta b=0 \tag{93}$$

It follows that for $x=b$ we have the relations

$$\frac{f_{y_1}'}{F_{y_1'}'}=\frac{f_{y_2}'}{F_{y_2'}'}=\frac{f_x'+f_{y_1}'y_1'+f_{y_2}'y_2'}{F} \tag{94}$$

which play the role of the transversality conditions in this problem and make it possible to determine all the necessary parameters.

Let us discuss the geometric significance of condition (94). Consider an arbitrary point (x, y_1, y_2) in the space. The values f_x', f_{y_1}', f_{y_2}' of the derivatives of $f(x, y_1, y_2)$ at this point (or, more precisely, the ratio $f_x':f_{y_1}':f_{y_2}'$ of the derivatives) specify the direction of the normal to the tangent plane drawn through the point (x, y_1, y_2) to the corresponding surface $f(x, y_1, y_2)=\text{const}$ passing through that point. Two relations (94) determine the values of the derivatives y_1', y_2' at that point. In the general case there can be one or several pairs of such values y_1', y_2' which thus specify a descrete set of directions in the space satisfying transversality conditions (94), the corresponding direction cosines (relative to the axes x, y_1, y_2) being in the ratio $1:y_1':y_2'$.

Now consider the case of a functional dependent on a curve (L) lying in the xy-plane which is represented by parametric equa-

tions (54). Let the functional have the form

$$I\{(L)\} = \int_{t_{\text{initial}}}^{t_{\text{terminal}}} \Phi(x, y)\sqrt{\dot{x}^2+\dot{y}^2}\,dt \tag{95}$$

Suppose that the coordinates of one of the end points of the curve (L) satisfy the equation

$$f(x, y) = 0 \tag{96}$$

We can interpret functional (95) as being defined for lines in the three-dimensional space t, x, y and consider equation (96) as a constraint of type (92). Then the application of equalities (94) leads to the transversality condition

$$f'_x : \frac{\Phi\dot{x}}{\sqrt{\dot{x}^2+\dot{y}^2}} = f'_y : \frac{\Phi\dot{y}}{\sqrt{\dot{x}^2+\dot{y}^2}}, \quad \text{i.e.} \quad \frac{\dot{x}}{f'_x} = \frac{\dot{y}}{f'_y}$$

But the latter relation is nothing but the orthogonality condition for the extremal $x = x(t)$, $y = y(t)$ and the curve $f(x, y) = 0$ in the xy-plane. Hence, for functional (95) the transversality condition turns into the orthogonality condition. It can be proved that (95) is the most general form of a functional for which the transversality condition reduces, at each point of the xy-plane, to orthogonality. This result is analogously extended to functionals dependent on curves in a Euclidean space of arbitrary dimension.

The result derived above can be given a visual interpretation in terms of geometrical optics. Suppose that a ray of light propagates in an isotropic medium (which may be nonhomogeneous in the general case). For simplicity, let us confine ourselves to the two-dimensional case and assume that the speed of light is equal to $c(x, y)$ at each point. Putting $\Phi = \frac{1}{c}$ in functional (95) we obtain the time taken for the ray of light to trace a given path (L) (why?). But the wave theory of light states that to the real rays of light there correspond stationary values of time, that is these rays go along extremals of the functional under consideration. If the rays of light issue from a point in all directions, the line (or surface, in the spatial case) formed of the end points of the rays at each time moment is called the **wave front**. The functional assumes stationary values at the points of this line, and, consequently, in this case the above assertion concerning the transversality condition for functionals of type (95) means that the rays of light are at each point perpendicular to the wave front.

Now let us proceed to investigate the case of two constraints of the form

$$f_1(x, y_1, y_2) = 0, \quad f_2(x, y_1, y_2) = 0 \tag{97}$$

imposed on the curve at the end point $x=b$. Then it is readily seen that the relations

$$f'_{ix}|_b\,\delta b+f'_{iy_1}|_b\,(\delta y_1|_b+y'_1|_b\,\delta b)+f'_{iy_2}|_b\,(\delta y_2|_b+y'_2|_b\,\delta b)=0 \qquad (i=1,2)$$

are fulfilled and imply equality (93). It follows (why?) that for $x=b$ we have the relation

$$\begin{vmatrix} F'_{y'_1} & F'_{y'_2} & F \\ f'_{1y_1} & f'_{1y_2} & f'_{1x}+f'_{1y_1}y'_1+f'_{1y_2}y'_2 \\ f'_{2y_1} & f'_{2y_2} & f'_{2x}+f'_{2y_1}y'_1+f'_{2y_2}y'_2 \end{vmatrix}=0 \qquad (98)$$

which serves as the transversality condition in this case.

Now let us consider the geometrical meaning of transversality condition (98). Take an arbitrary point M in the xy_1y_2-space. Let a curve described by equations of type (97) pass through this point. The direction of the tangent line k to this curve at the point M is determined by the values of the derivatives of the functions f_1 and f_2 entering into determinant (98). On the other hand, the derivatives y'_1, y'_2 specify the direction of the tangent line l to the extremal whose direction cosines are in the ratio $1:y'_1:y'_2$. Consequently, given a direction k at a point M, transversality condition (98) specifies a set of directions l which form one or several cones with vertices at M.

Functionals dependent on an arbitrary number n of unknown functions are investigated in a similar way. Let one of the ends of the sought-for line in an $(n+1)$-dimensional space with coordinates x, y_1, y_2, ..., y_n be allowed to move in a k-dimensional manifold (S). Then, at each point M, depending on the direction of the k-dimensional tangent hyperplane (P) to the manifold (S), the corresponding transversality conditions determine one or several $(n+1-k)$-dimensional conical hypersurfaces with vertices at M formed of the directions satisfying the transversality conditions.

The same approach is used for investigating functionals (43) involving derivatives of higher order. By the way, putting $y'=z$ we can reduce the problem to the corresponding conditional extremum problem for the functional

$$\int_a^b F(x,\ y,\ z,\ z')\,dx$$

with the differential constraint $y'-z=0$. By virtue of Sec. 12, this leads to transversality conditions of type (94) or (98) corresponding to the function $F^*=F(x,\ y,\ z,\ z')-r(x)\,(y'-z)$. Let the reader consider the case when in the original problem with functional (43) there is a constraint of the form $f(x,\ y)=0$ at one

of the ends and deduce the conditions

$$F'_{y'}=0,\ f'_yF-\left(F'_{y'}-\frac{d}{dx}F'_{y'}\right)(f'_x+f'_yy')=0$$

for this end point.

18. Transversality Condition for Functions of Several Variables. Here we shall consider a variational problem for a functional of a function of several independent variables in which the extremizing surface has a moving boundary. Let the domain (G) corresponding to a functional of form (55) vary with the surface $z=z(x, y)$ $((x, y)\in(G))$ for which the values of the functional are compared. Thus, the shape of the domain (G) is not fixed beforehand, and we only require that the contour bounding the variable surface $z=z(x, y)$ should lie on a given surface (S) with equation

$$f(x, y, z)=0 \qquad (99)$$

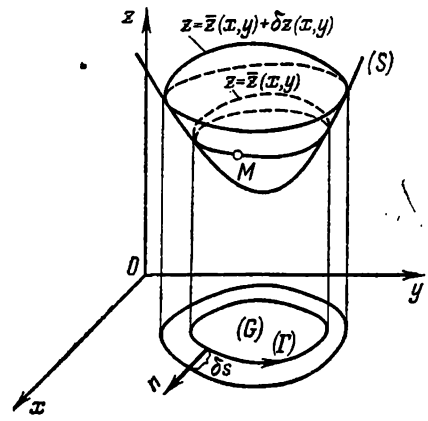


Fig. 89

The points of this contour have the coordinates $(x, y, z(x, y))$ $((x, y)\in(\Gamma)$ where (Γ) is the boundary of $(G))$, and consequently, the requirement that this contour lies on (S) reduces to the condition $f(x, y, z(x, y))=0$, $(x, y)\in(\Gamma)$. Reasoning as in Sec. 15, we conclude that the function $z=\bar{z}(x, y)$ extremizing the functional satisfies the Euler-Ostrogradsky equation (59). Furthermore, taking into account the variability of the domain (G) and integrating by parts, like in formula (58), we obtain, by virtue of equation (59), the following expression for the variation of functional (55):

$$\delta I\{\bar{z},\ \delta z\}=\int\limits_{(\Gamma)}F\delta s\,d\Gamma+\oint\limits_{(\Gamma)}\delta z\,(F'_p\,dy-F'_q\,dx) \qquad (100)$$

Here δs designates the width (which is not constant in the general case and is understood as an algebraic quantity that may assume both positive and negative values) of the annular region (see Fig. 89) which is added ("algebraically") to the domain (G) when it varies. The first integral in (100) expresses the contribution made to the variation of the functional by the variation of the domain (G), and the second integral appears after the integration by parts and is taken in the positive direction of the contour (Γ).

Let us denote by $\mathbf{n}$ the unit vector of outer normal to the contour (Γ) (this unit vector varies as the moving point traces the contour (Γ)). Then if we take an arbitrary point belonging to (Γ) and then pass, along $\mathbf{n}$, to the corresponding point lying on the other closed contour bounding the above-mentioned annular region (see Fig. 89), the coordinates of the point receive the increments

$$\delta x = n_x \delta s \quad \text{and} \quad \delta y = n_y \delta s$$

Therefore, by virtue of (99), it follows that

$$f'_x n_x \delta s + f'_y n_y \delta s + f'_z \left(\delta z + \frac{\partial z}{\partial x} n_x \delta s + \frac{\partial z}{\partial y} n_y \delta s \right) = 0$$

Let us express δz from the latter relation, substitute it into the second integral in formula (100) and then transform this integral into the corresponding line integral with respect to arc length by applying the formulas $dx = -n_y d\Gamma$ and $dy = n_x d\Gamma$. This results (check it up!) in

$$\delta I = \int_{(\Gamma)} \left[F - (F'_p n_x + F'_q n_y) \left(\frac{f'_x n_x + f'_y n_y}{f'_z} + \frac{\partial z}{\partial x} n_x + \frac{\partial z}{\partial y} n_y \right) \right] \delta s \, d\Gamma$$

Now we can use necessary condition (25) for extremum. Applying (25) and taking into account that δs is an arbitrary quantity we conclude that the expression in the square brackets under the integral sign must be equal to zero at each point of the contour (Γ). The relation thus obtained is the transversality condition for the problem in question. At each point M (whose coordinates satisfy a condition of type (99)) this transversality condition connects the values of the derivatives $\frac{\partial z}{\partial x}$ and $\frac{\partial z}{\partial y}$, the relationship between these derivatives being dependent on the position of the tangent plane to (S) at the point M.

By analogy with (83), we can take a more general functional

$$\int_{(G)} F\left(x, y, z, \frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right) dG + \int_{(\Gamma)} \varphi(x, y, z) \, d\Gamma$$

dependent on a function of two arguments. Let us consider an analogous variational problem for this functional with fixed domain (G). Performing transformations similar to the above we can substitute, into the integral of the type of the second term in formula (100) which appears in these calculations, the expressions $dy = \cos(\widehat{\mathbf{n}, x}) \, d\Gamma$, $dx = -\cos(\widehat{\mathbf{n}, y}) \, d\Gamma$ and then obtain, by the arbitrariness of $\delta z|_{(\Gamma)}$, the *natural boundary condition*

$$F'_p \cos(\widehat{\mathbf{n}, x}) + F'_q \cos(\widehat{\mathbf{n}, y}) + \varphi'_z = 0 \qquad (101)$$

This result is similarly extended to functionals involving functions of an arbitrary number of independent variables.

19. Unilateral Constraints. Let us come back to functional (18) and suppose that the variables x and y are subject to a finite constraint of the form

$$f(x, y) \geqslant 0$$

A constraint of this type involving an inequality is referred to as **unilateral**, the terms **one-sided** or **non-restricting** (constraint) also being used in this case. (In the foregoing problems we considered subsidiary conditions expressed by equalities which are called **bilateral** or **two-sided** constraints.)

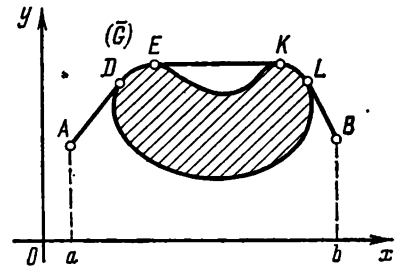


Fig. 90

The condition $f(x, y, z) \geqslant 0$ means that the graphs of the functions $y(x)$ for which the values of the functional are compared lie in a given closed (that is including its boundary) domain $(\bar{G})$ in the xy-plane. A typical example of this kind is the problem of finding the shortest path lying within a given domain $(\bar{G})$ and connecting two given points (see Fig. 90 where the domain $(\bar{G})$ is the closure of the exterior of the shaded region). This problem reduces to minimizing the functional

$$I\{y\} = \int_a^b \sqrt{1+y'^2}\, dx \qquad (102)$$

The graph of the function $\bar{y}(x)$ extremizing functional (102) consists of parts lying in the interior of $(\bar{G})$ (these are the parts AD, EK and LB in Fig. 90) and of parts belonging to the boundary of $(\bar{G})$ (see DE and KL). The argument similar to that in Sec. 13 shows that the former parts are extremals, i. e. integral curves of Euler's equation, while, generally speaking, the latter parts are not extremals. Therefore, in the general case, the sought-for extremal value $I\{\bar{y}\}$ is not a stationary value of the functional and can be compared to an extremum of a function attained on the boundary of its domain of definition (e.g. see [107], Secs. IV.19 and XII.11). The necessary condition of extremum for the parts of the second type depends on the type of the extremum and on the admissible direction in which the curves (for which the values of the functional are compared) can move in accordance with the unilateral constraint. For instance, in the above minimum problem

the parts DE and KL shown in Fig. 90 can only be moved in the direction of increase of the coordinate y (in these cases we speak of **one-sided variations** of the extremizing curve). Giving the function $y=y(x)$ the admissible one-sided variations corresponding to these parts and integrating by parts we arrive at the ordinary expression for the variation: $\delta I=\int\left(F'_y-\frac{d}{dx}F'_{y'}\right)\delta y\,dx$. But now the necessary condition for extremum is that we must have $\delta I\geqslant 0$ for every $\delta y\geqslant 0$. It follows that, on these parts, we must have

$$F'_y-\frac{d}{dx}F'_{y'}\geqslant 0 \tag{103}$$

(let the reader carefully examine this consideration). Inequality (103) expresses the desired necessary condition. If, instead of the above minimum problem for (102), we consider a maximum of a functional in the domain $(\bar{G})$ or if the admissible direction of variation of the parts of the curve lying on the boundary of the domain is reversed the sign of inequality in (103) must be changed to the opposite.

Taking functional (102) we see that condition (103) reduces to the inequality $y''\leqslant 0$ (check it up!), and thus we draw the obvious conclusion that in this case the sought-for extremal can only go along those parts of the boundary of the shaded domain in Fig. 90 where it is convex upward (why?).

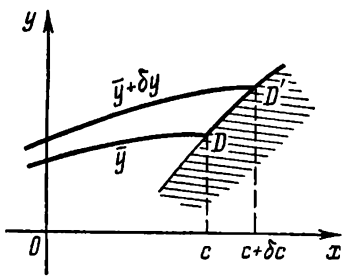


Fig. 91

Now let us establish the necessary condition for a point at which the graph of the function $\bar{y}(x)$ passes from the interior of the domain $(\bar{G})$ to its boundary or in the opposite direction, e.g. the point D which we consider below and the points E, K and L in Fig. 90. To this end, we consider the values of δI corresponding to an arbitrary (admissible) variation δy in the vicinity of that point (see Fig. 91). By analogy with (91), applying the mean value theorem to the integral along DD' (which enters into the variation with the minus sign) we obtain the expression

$$\delta I=F'_{y'}|_{x=c}k\delta c+(F-y'F'_{y'})|_{x=c}\delta c-F(c,\ \bar{y}(c),\ k)\,\delta c$$

where k is the slope of the tangent to the boundary of the domain $(\bar{G})$ at the point D (where the curve $y=\bar{y}(x)$ joins the boundary). The value of the quantity δc being of arbitrary sign, we thus derive the necessary condition which must be fulfilled

at every point of the type of D:

$$F'_{y'}(c, \bar{y}(c), \bar{y}'(c))\,[k-\bar{y}'(c)]+ \\ +F(c, \bar{y}(c), \bar{y}'(c))-F(c, \bar{y}(c), k)=0 \qquad (104)$$

This condition is sure to be fulfilled if $\bar{y}'(c)=k$, that is if the graph of $\bar{y}(x)$ has no corner at the point under consideration. There is a simple case when the absence of the corner point is necessary, namely when the derivative $F''_{y'y'}$ at the point D does not change its sign for all y' and is not identically equal to zero. Indeed, the left-hand side of (104) is equal (check it up!) to

$$-\int_{\bar{y}'(c)}^{k} dr \int_{\bar{y}'(c)}^{r} F''_{y'y'}(c, \bar{y}(c), q)\,dq$$

and hence, under the above condition on $F''_{y'y'}$, we always have $\bar{y}'(c)=k$ at the point D. In particular, this is the case when the function $F''_{y'y'}(x, y, y')$ is continuous and has no zeros.

In this connection we note that if $F''_{y'y'}(x, y, y')$ turns into zero for certain values of x, y, y' there arise considerable difficulties in extrema problems for functional (18) for domain in which these zero values may appear. In fact, we see that, for the values of the variables corresponding to a zero of $F''_{y'y'}$, Euler's equation (37) degenerates, that is its order is reduced, and as is known, in these cases the initial conditions may specify more than one solution or none, etc. Let the reader show that for functionals of type (46) involving an arbitrary number n of functions y_k, the values of the variables for which the determinant $\det\left(F''_{y'_k y'_l}\right)$ of the nth order vanishes play an analogous role.

Problems with unilateral constraints for curves in the three-dimensional space are treated in [80], § 79.

20. Extremals with Corners. There are many problems of the calculus of variations in which the extremum is attained on piecewise smooth curves, that is on extremals with corner points whose equations involve continuous functions with derivatives having discontinuities at separate points.

Let us begin with a problem in which the fact that the sought-for curve has a corner point is directly implied by the formulation of the problem. We shall consider the following **reflection-of-extremal problem:** it is required to determine an extremum of a functional of the form

$$I\{y_1, y_2\}=\int_{a_1}^{b} F_1(x, y_1, y'_1)+\int_{a_2}^{b} F_2(x, y_2, y'_2)\,dx$$

where F_1 and F_2 are different functions and the unknown functions $y_1(x)$ and $y_2(x)$ satisfy the boundary conditions

$$y_1(a_1) = y_{10}, \quad y_2(a_2) = y_{20}, \quad y_1(b) = y_2(b) \tag{105}$$

(see Fig. 92). Considerations analogous to those in Sec. 13 indicate that if a pair of functions $\bar{y}_1(x)$, $\bar{y}_2(x)$ extremizes the functional, the function $\bar{y}_1(x)$ is a solution of Euler's equation corresponding to the function F_1 and $\bar{y}_2(x)$ satisfies Euler's equation corresponding to the function F_2. (In particular, if $F_1 \equiv F_2$ we obtain a piecewise smooth extremal whose both parts satisfy the same Euler equation.) As in Sec. 13, we obtain the expression

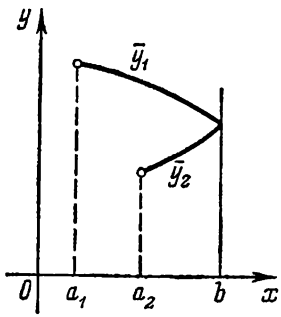


Fig. 92

$$\delta I\{\bar{y}_1, \bar{y}_2, \delta y_1, \delta y_2\} = F'_{1y'_1}(b, \bar{y}_1(b), \bar{y}'_1(b))\,\delta y_1(b) + F'_{2y'_2}(b, \bar{y}_2(b), \bar{y}'_2(b))\,\delta y_2(b)$$

Consequently, the general necessary condition for extremum (equality (25)) and conditions (105) imply the following necessary condition that must hold at the point of reflection:

$$F'_{1y'_1}(b, y_1(b), y'_1(b)) + F'_{2y'_2}(b, y_2(b), y'_2(b)) = 0$$

If, under the same conditions, we have $a_2 > b$, it appears more natural to speak about the **refraction of the extremal.**

If the line on which the point of reflection lies is not vertical, as in the above problem, but is an arbitrary curve we can apply the argument used in deriving formula (90) and thus obtain the more general reflection condition

$$(k - l_1)F'_{1y'}(x, y, l_1) + F_1(x, y, l_1) + (k - l_2)F'_{2y'}(x, y, l_2) + F_2(x, y, l_2) = 0 \tag{106}$$

where k, l_1 and l_2 are, respectively, the slopes of the tangents drawn through the point of reflection to the line of reflection, of the "incident" extremal and of the "reflected" extremal.

Let us now investigate the case when the presence of a corner point of the extremal is not meant in the original formulation of the problem and appears in a natural way in the process of solution. Suppose that functional (18) with boundary conditions (26) is considered as being defined not only for smooth functions belonging to $C_1[a, b]$ but also for piecewise smooth functions $y(x)$. Then, like in Sec. 13, we can draw the conclusion that if a piece-

wise smooth function $y(x)$ gives at least a weak extremum to the functional, it satisfies Euler's equation on the intervals lying between the corner points (let the reader formulate the precise definition of a weak extremum for this case by analogy with the foregoing sections). If M is a corner point whose abscissa is x, we can write functional (18) in the form

$$\int_a^b F\,dx = \int_a^x F\,dx + \int_b^x (-F)dx$$

and take as a line of refraction an arbitrary curve passing through the point M. Then we can apply condition (106) which thus must be an identity with respect to k. Putting $F_1 = F$, $F_2 = -F$ and equating the coefficient in k and the constant term in the left-hand side of (106) to zero we see (check it up!) that the values of

$$F'_{y'}(x, y, y') \text{ and } F(x, y, y') - y'F'_{y'}(x, y, y') \qquad (107)$$

must vary continuously when the variable point passes through the corner point although the derivative y' has a jump at this point. In other words, *each of the functions* $F'_{y'}$ *and* $F - y'F'_{y'}$ *must assume the same limiting values on both sides of the corner point.* This is known as the **Weierstrass-Erdmann conditions.** The first condition concerning $F'_{y'}$ implies that a corner can only be at the points (x, y) where the derivative $F''_{y'y'}(x, y, y')$ vanishes (because, if otherwise, the function $F'_{y'}(x, y, l)$ is monotone with respect to l and therefore the equality $F'_{y'}(x, y, l_1) = F'_{y'}(x, y, l_2)$ implies that $l_1 = l_2$, i.e. there is no corner at the point (x, y)). It can be proved that the other condition (concerning the second function (107)) implies that, as y' varies, the expression $F''_{y'y'}(x, y, y')$, considered for the coordinates (x, y) of the corner point, must turn into zero at least twice.

It can readily be shown that the Weierstrass-Erdmann conditions for functionals of several functions (for instance, functionals of form (46)) are reduced to the requirements that the expressions $F'_{y'_1}$, $F'_{y'_2}$ and $F - y'_1F'_{y'_1} - y'_2F'_{y'_2}$ must be continuous at the corner point.

Let us discuss a particular case of the third problem of Sec. 1 in which there appears a solution with corners. In Sec. 8 we showed that the trace of the sought-for surface of revolution in the xy-plane is obtained from the graph of the function $y = \cosh x$ by the transformation of similitude which makes the transformed graph pass through the points of intersection of the wire rings with the xy-plane. In Fig. 93 the part of the section of the surface by the xy-plane lying in the first quadrant is shown in heavy line, the corresponding point of intersection being M_1. The family of curves obtai-

ned from $y=\cosh x$ by means of all the similarity transformations entirely covers an angle $\frac{y}{x} \geqslant k=\tan\alpha$ (let the reader derive the transcendental equation for k and find from it the approximate numerical value $k=1.51$). But if the point of intersection is inside the angle $\frac{y}{x}<k$ there is no smooth extremal passing through it (see the point M_2 shown in Fig. 93). Therefore, it appears natural to pose the problem of determining the shape of the soap film when the distance between the rings is sufficiently large, that is when the points of intersection of the rings with the xy-plane occupy the positions of the type of M_2. On the basis of physical intuition we can expect that the film will split into two separate disks spanned by the two rings because its area must be as small as possible. We can also imagine that the centres of the disks are joint by the "soap cylinder" of zero width going along the axis of revolution, which does not affect the total area. Then the trace of this surface in the first quadrant of the xy-plane is the broken line ONM_2. The Weierstrass-Erdmann conditions are obviously fulfilled for all the points of the x-axis because all the expressions involved vanish at these points. Strictly speaking, we cannot speak about the Weierstrass-Erdmann conditions for the portion NM_2 of the extremal on which y cannot be expressed as a single-valued function of x. But nevertheless the above intuitive considerations can be justified rigorously because we can transform the problem to an equivalent parametric variational problem of the type considered in Sec. 10 with a functional of the form $\int F(x, y, \dot{x}, \dot{y})\,dt$, and for this functional it can easily be proved that the Weierstrass-Erdmann conditions are reduced to the requirement that derivatives $F'_{\dot{x}}$ and $F'_{\dot{y}}$ are continuous.

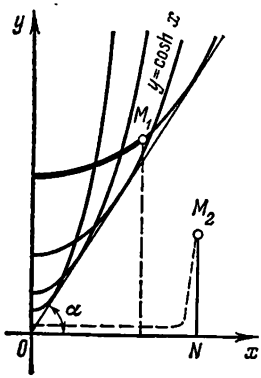


Fig. 93

If we consider only smooth curves for the point M_2 we can take an approximate smooth solution generating a surface whose area is as close as desired to the minimum area (see a curve of this type shown in dotted line in Fig. 93), but the exact minimum area cannot be obtained in this way. Consequently, this problem, which has been originally posed for the space C_1, has no solution in C_1 and, to obtain the solution, we must extend the space in an appropriate manner. In these cases we usually speak about the **generalized solution** of the original problem.

§ 2. SECOND VARIATION AND SUFFICIENT CONDITIONS FOR EXTREMUM

We have already spoken of a close analogy between extrema problems of the differential calculus and the calculus of variations (Secs. 1.4 and 5). Therefore, it appears natural that the sufficient conditions for variational extremum must involve a notion analogous to the second differential which plays an essential role in the theory of extrema of functions (e.g. see [107], Secs. IV.18 and XII.7). This new notion is the *second variation* of a functional which will be investigated in § 2.

1. Variations of Higher Order. The increment of a functional can be replaced by its variation in a more precise manner than in Sec. 1.4 if we take into account the terms of higher orders of smallness (when the increment of a function is replaced by its differential the analogous role is played by Taylor's expansion of the function; e.g. see [107], formula (IV.62)). This leads to the series expansion

$$\mathrm{I}\{y+\delta y\}-\mathrm{I}\{y\}=\delta I\{y,\ \delta y\}+$$
$$+\frac{1}{2!}\delta^2 I\{y,\ \delta y\}+\frac{1}{3!}\delta^3 I\{y,\ \delta y\}+\frac{1}{n!}\delta^n I\{y,\ \delta y\}+\ldots \qquad (1)$$

which is completely analogous to Taylor's series for a function. Here $\delta^2 I$, $\delta^3 I$, ..., $\delta^n I$, ... are, respectively, the **second, third**, ..., ***n*th**, etc. **variations** of the functional I. Each of them is the variation of the preceding variation, that is $\delta^2 I=\delta(\delta I)$, etc. where the function δy is regarded as independent of y, that is remains invariable. The kth variation ($k=1, 2, \ldots$) possesses the homogeneity property with degree k relative to δy:

$$\delta^n I\{y,\ C\delta y\}=C^n\delta^n I\{y,\ \delta y\} \qquad (C=\text{const})$$

This means that the subsequent terms on the right-hand side of (1) are, respectively, of the first, second, third, etc. order of smallness with respect to δy (more precisely, with respect to the norm $\|\delta y\|$ taken in accordance with the space in question for which the functional is considered; see Sec. 1.3) provided that they do not vanish identically.

In § 1 we used only the first term of series (1). In § 2 we shall also use the second term, that is we shall take the more precise formula

$$I\{y+\delta y\}-I\{y\}=\delta I\{y,\ \delta y\}+\frac{1}{2!}\delta^2 I\{y,\ \delta y\}+O(\|\delta y\|^3) \qquad (2)$$

In concrete examples expansions of form (1) (and therefore of form (2) as well) are obtained by means of the ordinary Taylor

series. For instance, taking functional (1.18) we obtain the expression

$$I\{y+\Delta y\}=\int_a^b F(x,\ y+\delta y,\ y'+\delta y')\,dx=$$

$$=\int_a^b\Big[F+(F'_y\delta y+F'_{y'}\delta y')+$$

$$+\frac{1}{2!}(F''_{yy}\delta y^2+2F''_{yy'}\delta y\delta y'+F''_{y'y'}\delta y'^2)+\ldots\Big]\,dx$$

where F, F'_y, etc. are taken for the corresponding values of x, y and y'. This relation yields formula (1.19) for the first variation and the formula

$$\delta^2 I\{y,\ \delta y\}=\int_a^b (F''_{yy}\delta y^2+2F''_{yy'}\delta y\,\delta y'+F''_{y'y'}\delta y'^2)\,dx \qquad (3)$$

for the second variation (here, of course, $\delta y^2=(\delta y)^2$ and $\delta y'^2=$ $=(\delta y')^2$).

For a fixed function $y(x)$, the expression $\delta^2 I\{y,\ \delta y\}$ is a **quadratic functional** with respect to δy. In the general case the notion of a quadratic functional $f(y)$ defined in an arbitrary linear space (R) is introduced in the following way. We first introduce the notion of a **bilinear functional** $\varphi(y_1,\ y_2)$ defined for all $y_1\in(R)$, $y_2\in(R)$ which is a linear functional of each of these elements y_1 and y_2 when the other element is kept fixed. After that we put

$$f(y)=\varphi(y,\ y)$$

and thus define the corresponding quadratic functional $f(y)$. If (R) is a normal linear space, a quadratic functional $f(y)$ defined in it is said to be **bounded** if it satisfies the inequality

$$|f(y)|\leqslant K\|y\|^2$$

The least value of K for which this relation holds (for all y) is referred to as the **norm** of the quadratic functional. In variational problems the space (R) in which a quadratic functional is considered is usually such that the functional is bounded.

The notions of a **multilinear functional** $\varphi(y_1,\ y_2,\ \ldots,\ y_n)$ and the corresponding functional $f(y)=\varphi(y,\ y,\ \ldots,\ y)$ are defined in the same manner. The terms on the right-hand side of (1) are nothing but functionals of that type relative to the variation δy.

Formula (1) can be derived by means of Taylor's formula for functions of one independent variable in just the same way as Taylor's formula for functions of several variables (e.g. see [107], Sec. XII.6). To this end, we fix y and δy and consider the function $I\{y+\lambda\delta y\}$ of the scalar parameter λ which is expanded

in ordinary Taylor's series in powers of λ:

$$I\{y+\lambda\delta y\} = I\{y\}\left[\frac{d}{d\lambda} I\{y+\lambda\delta y\}\right]_{\lambda=0}\lambda + \\ + \frac{1}{2!}\left[\frac{d^2}{d\lambda^2} I\{y+\lambda\delta y\}\right]_{\lambda=0}\lambda^2 + \ldots \quad (4)$$

The expressions in the square brackets are obviously the variations of the corresponding orders. Putting $\lambda = 1$ in (4) we obtain formula (1).

There is another method of deducing formula (1) which is analogous to the method used in Sec. 1.12 for deriving the expression of function (64). Namely, replacing approximately the function $y(x)$ by a finite set of its values we thus replace the functional $I\{y\}$ by the function $I(y_0, y_1, \ldots, y_n)$ of n scalar variables. Now we can expand the increment

$$\Delta I = I(y_0+\delta y_0,\ y_1+\delta y_1,\ \ldots,\ y_n+\delta y_n) - I(y_0, y_1, \ldots, y_n)$$

of this function in powers of the increments of the variables and then return to the functionals and thus obtain expansion (1). This argument clarifies the role of the second variation in the theory of extrema of functionals.

2. Expressing Sufficient Conditions for Extremum in Terms of Second Variation. Let a functional $I\{y\}$ assume a stationary value for a function $y=\bar{y}$, that is $\delta I\{\bar{y}, \delta y\} \equiv 0$; in other words, let necessary condition (1.25) for extremum be fulfilled. Then the substitution of $y=\bar{y}$ into the right-hand side of (2) turns the first term into zero, and therefore the second term becomes dominant. Therefore, reasoning as in the case of an extremum of a function of several variables (e.g. see [107], Sec. XII.7) we draw the following conclusions:

If $\delta^2 I > 0$ for every variation δy (not equal, of course, identically to zero because this results in $\delta^2 I = 0$), *the function $y = \bar{y}(x)$ minimizes the functional $I\{y\}$;*

if $\delta^2 I < 0$ for every δy the function $y=\bar{y}(x)$ maximizes the functional $I\{y\}$;

if $\delta^2 I$ can assume values of both signs, the functional $I(y)$ has a minimax for $y=\bar{y}(x)$, and thus, there is no extremum.

The only ambiguous case, when the behaviour of $\delta^2 I$ does not enable us to say whether there exists an extremum, appears when $\delta^2 I$ does not change its sign but may turn into zero (or, in particular, is identically equal to zero as functional dependent on δy). Here we shall not treat this more complicated case.

A more thorough investigation shows that in the general case the sufficient condition for minimum is written in the form

$\delta^2 I\{\bar{y}, \delta y\} \geqslant C\|\delta y\|^2$ $(C = \text{const} > 0)$. For functional spaces this condition is stronger than the requirement $\delta^2 I > 0$. But usually, for ordinary variational problems, this distinction is inessential.

It is desirable to write the conditions obtained above in the form of equivalent conditions imposed directly on the unknown function $\bar{y}(x)$, as was done for necessary condition (1.25). This transformation of sufficient conditions is carried out differently depending on the type of the functional under consideration. Here we shall demonstrate these techniques for some classes of problems.

It should be noted that the type of the extremum (weak or strong, etc.; see Sec. 1.3) depends on the choice of the space in which the functional $I\{y\}$ and expansions (1) and (2) are considered.

3. Legendre's Necessary Conditions. Let us consider functional (1.18) with boundary conditions (1.26) or more general conditions mentioned in Sec. 1.13. We showed that the second variation of this functional is expressed by formula (3). It can be easily proved that

if $F''_{y'y'}(x, \bar{y}(x), \bar{y}'(x))$ *takes on positive values for some* x *belonging to the interval* $a \leqslant x \leqslant b$ (not necessarily for all $x \in [a, b]$), *then* $\delta^2 I\{\bar{y}, \delta y\}$ *also assumes positive values for some* δy;

if $F''_{y'y'}(x, \bar{y}(x), \bar{y}'(x))$ *assumes for some* $x \in [a, b]$ *negative values, the second variation* $\delta^2 I\{\bar{y}, \delta y\}$ *also takes on negative values for some* δy.

Indeed, let us take the first case and substitute $y = \bar{y}(x)$ into the coefficients of the quadratic form under the sign of integration in (3) and denote the values of the coefficients thus obtained by $\varphi(x)$, $\psi(x)$ and $\chi(x)$, respectively. Suppose that $\chi(x_0) > 0$ where $a < x_0 < b$ (such a value x_0 is sure to exist because, since all the functions in question and their derivatives are considered to be continuous, the assumption that $\chi(x) \leqslant 0$ for $a < x < b$ implies that $\chi(x) \leqslant 0$ for all $x \in [a, b]$ as well). Now let us take, as δy, the function whose graph is shown in heavy line in Fig. 94. (If we want to limit ourselves to functions belonging to C_1 we can change the shape of the graph in small neighbourhoods of its corner points in such a way that the resultant graph is smooth and as close as desired to that in Fig. 94.) For the variation δy

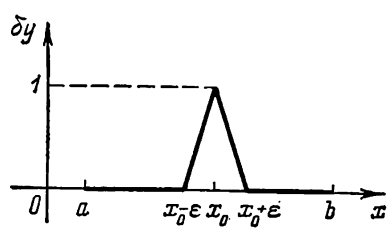


Fig. 94

of this particular form we obtain the expression

$$\delta^2 I\{\bar{y},\ \delta y\} = \int_{x_0-\varepsilon}^{x_0+\varepsilon} \varphi(x)\,\delta y^2\,dx + \\ + 2\int_{x_0-\varepsilon}^{x_0+\varepsilon} \psi(x)\,\delta y\delta y'\,dx + \int_{x_0-\varepsilon}^{x_0+\varepsilon} \chi(x)\,\delta y'^2\,dx \qquad (5)$$

If $\varepsilon > 0$ is sufficiently small we can replace the coefficients $\varphi(x)$, $\psi(x)$ and $\chi(x)$ under the integral sign in (5) by their values at the point $x = x_0$ and thus receive the approximate values

$$\varphi(x_0)\cdot\frac{2}{3}\varepsilon,\quad \psi(x_0)\cdot 0 \quad \text{and} \quad \chi(x_0)\cdot\frac{2}{\varepsilon}$$

of the three terms entering into the right-hand side of (5). Consequently, for sufficiently small values of $\varepsilon > 0$ the whole sum (5) is positive, which is what we set out to prove. The second case is investigated in the same manner.

Comparing this result with the assertions established in Sec. 2 we arrive at **Legendre's necessary conditions:** *if, under certain boundary conditions, functional* (1.18) *has a minimum* (at least weak minimum) *for* $y = \bar{y}(x)$, *then* $F''_{y'y'}(x, \bar{y}(x), \bar{y}'(x)) \geqslant 0$ $(a \leqslant x \leqslant b)$; *if it has a maximum then* $F''_{y'y'}(x, \bar{y}(x), \bar{y}'(x)) \leqslant 0$ $(a \leqslant x \leqslant b)$. Incidentally we have obtained the following sufficient condition for minimax: *if the function* $F''_{y'y'}(x, \bar{y}(x), \bar{y}'(x))$ *assumes values of both signs* (*for* $a \leqslant x \leqslant b$) *along an extremal* $\bar{y}(x)$, *functional* (1.18) *has a minimax for* $y = \bar{y}(x)$. (Let the reader carefully examine all these assertions.)

In the general case Legendre's necessary conditions for extremum are not sufficient (see Sec. 5).

Functionals of several functions (for instance, functionals of form (1.46) dependent on two functions) are investigated in like manner. Expansions (1) and (2) are valid for these functionals as well, but in this case the quantities y and δy are vector functions. For the second variation we obtain, instead of (3), the formula

$$\delta^2 I\{\mathbf{y},\ \delta\mathbf{y}\} = \int_a^b \left(\sum_{i,\,j} F''_{y_i y_j}\delta y_i \delta y_j + \sum_{i,\,j} F''_{y_i y'_j}\delta y_i \delta y'_j + \sum_{i,\,j} F''_{y'_i y'_j}\delta y'_i \delta y'_j\right) dx$$

where, for a functional of type (1.46), the indices i and j independently take on the values 1 and 2, and, for the case of n unknown functions, the values 1, 2, ..., n. Reasoning as above, we arrive at Legendre's necessary conditions which in this case state that if a vector function $\mathbf{y} = \bar{\mathbf{y}}(x)$ minimizes the functional

under consideration (taken with certain boundary conditions), the quadratic form

$$\sum_{i,\ j} F''_{y_i' y_j'}(x,\ \bar{\mathbf{y}}(x),\ \bar{\mathbf{y}}'(x))\,\xi_i\xi_j \tag{6}$$

cannot assume negative values for any x belonging to the interval $a \leqslant x \leqslant b$, that is it must be nonnegative, which means that all the eigenvalues of the matrix of the quadratic form are nonnegative; if the function $\mathbf{y}=\bar{\mathbf{y}}(x)$ maximizes the functional, quadratic form (6) must be nonpositive on the entire interval $a \leqslant x \leqslant b$; finally, if the matrix $(\mathbf{F}''_{y_i y_j})$ has eigenvalues of both signs (at least for different values of x), the functional has a minimax on the extremal $\mathbf{y}=\bar{\mathbf{y}}(x)$.

4. Quadratic Functional. Let us consider the functional

$$I\{y\} = \int_a^b [P(x)\,y'^2 + Q(x)\,y^2]\,dx \tag{7}$$

in which the function $y(x)$ satisfies the boundary conditions

$$y(a)=0,\quad y(b)=0 \tag{8}$$

We shall suppose that the coefficients $P(x)$ and $Q(x)$ are continuous (although it is sufficient to assume that they are finite). Besides, in what follows we shall always require that

$$P(x) > 0 \quad (a \leqslant x \leqslant b) \tag{9}$$

Euler's equation for quadratic functional (7) takes the form

$$(P(x)\,y')' - Q(x)\,y = 0 \tag{10}$$

and thus is a second-order homogeneous linear differential equation. In particular, the function $y \equiv 0$ always satisfies equation (10) and conditions (8), and hence functional (7) receives a stationary value for $y \equiv 0$, which, by the way, is implied directly by the conditions of the problem (why?). To find out whether this stationary value is extremal we shall prove the following proposition:

Let $y = Y(x)$ be the solution of equation (10) *satisfying the initial conditions*

$$y(a)=0,\quad y'(a)=C\neq 0 \tag{11}$$

If the function $Y(x)$ has no zeros for $a < x \leqslant b$, the quantity $I\{0\}=0$ is a minimum value of the functional. If $Y(x)$ has at least one zero on the interval $a < x < b$, $I\{0\}=0$ is a minimax value of the functional.

Initial values (11) are directly proportional to C, and therefore, if the constant C is varied, the function $Y(x)$ also varies

proportionally. Hence, its zeros are independent of C, and we can simply put $C=1$.

Let $\tilde{a}$ be the first zero of the function $Y(x)$ lying to the right of the point $x=a$. Then the point $x=\tilde{a}$ is said to be **conjugate** to the point $x=a$ (relative to equation (10) or to functional (7)). If there are no such zeros and the coefficients of equation (10) are defined for $a \leqslant x < \infty$ we can conditionally put $\tilde{a}=\infty$. Thus, the above proposition can be equivalently formulated as follows: *if $\tilde{a} > b$, functional $I\{y\}$ has a minimum for $y \equiv 0$, and if $\tilde{a} < b$, it has a minimax.* (It can be shown that if $\tilde{a}=b$ the functional has a nonstrict minimum for $y \equiv 0$.)

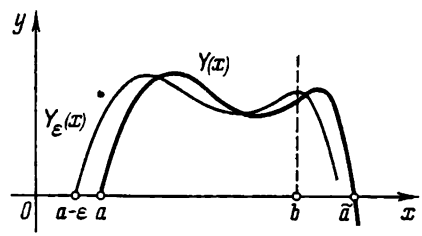


Fig. 95

We begin with proving the first part of the proposition. Let $\tilde{a} > b$; then the solution $Y_\varepsilon(x)$ of equation (10) satisfying the initial conditions $Y_\varepsilon(a-\varepsilon)=0$, $Y'_\varepsilon(a-\varepsilon)=C$ is different from zero on the whole interval $a \leqslant x \leqslant b$ for all sufficiently small $\varepsilon > 0$ (see Fig. 95) because, as is known (e.g. see [107], Sec. XV.28), the solution of a differential equation of type (10) and its derivatives are continuous functions of the initial values. By virtue of (8), we have the relation

$$\int_a^b d[v(x)\,y^2] = v(x)\,y^2\Big|_a^b = 0$$

for any function $v(x)$. Let us add this expression to the right-hand side of (7) and choose the function $v(x)$ such that the new integrand function

$$Py'^2 + Qy^2 + (vy^2)' = Py'^2 + 2vy'y + (v'+Q)y^2 = \\ = P\left(y'^2 + 2\frac{v}{p}y'y + \frac{v'+Q}{p}y^2\right)$$

is a complete square. For this requirement to hold we must have $\frac{v'+Q}{p} = \left(\frac{v}{p}\right)^2$, and it is readily verified that the function $v = -\frac{PY'_\varepsilon}{Y_\varepsilon}$ satisfies this equation (check it up!). This construction shows that $I\{y\} > 0$ for $y \not\equiv 0$, which is what we set out to prove. (Let the reader find out what part of this proof is based on the condition that $\tilde{a} > b$.)

Now we proceed to prove the second part of the proposition. Let $\tilde{a} < b$; then, introducing the function

$$w(x) = \begin{cases} Y(x) & \text{for } a \leqslant x \leqslant \tilde{a} \\ 0 & \text{for } \tilde{a} \leqslant x \leqslant b \end{cases}$$

we see that it is a piecewise smooth extremal (see Sec. 1.20). But the Weierstrass-Erdmann condition is not fulfilled at the corner point $x = \tilde{a}$ (check it up!), and therefore $I\{w\}$ is not an extremal value of the functional. Consequently, there exist functions $w_1(x)$ and $w_2(x)$ such that $I\{w_1\} > 0$ and $I\{w_2\} < 0$. Therefore, we have $I\{\lambda w_1\} > 0$ and $I\{\lambda w_2\} < 0$ for any $\lambda = \text{const} \neq 0$, which indicates that there is a minimax for $y \equiv 0$. The proof of the proposition has thus been completed.

Let us consider the particular case of constant coefficients $P > 0$ and Q. It is clear that if $Q \geqslant 0$, $I\{0\}$ is a minimum value of the functional. Suppose that $Q < 0$; then the solution of equation (10) satisfying initial condition (11) has the form $y = C\sqrt{\frac{P}{|Q|}} \sin \sqrt{\frac{|Q|}{P}}(x-a)$ (check it up!), and its first zero to the right of a is at the point $x = \tilde{a} = a + \pi\sqrt{\frac{P}{|Q|}}$. Hence, if the difference $b-a$ is less than $\pi\sqrt{\frac{P}{|Q|}}$ the functional under consideration achieves its minimum value for $y \equiv 0$, and if $b-a$ is greater than $\pi\sqrt{\frac{P}{|Q|}}$ it has a minimax.

The quadratic functional of the general form

$$I\{y\} = \int_a^b [P(x)\, y'^2 + R(x)\, y' y + Q(x)\, y^2]\, dx \qquad (12)$$

with boundary conditions (8) is reduced to form (7) if we apply integration by parts to the integral $\int_a^b R(x)\, y' y\, dx$, which results in

$$I\{y\} = \int_a^b \left[P(x)\, y'^2 + \left(Q(x) - \frac{1}{2} R'(x)\right) y^2\right] dx \qquad (13)$$

Euler's equations for functionals (12) and (13) coincide (let the reader check up all these properties). Consequently, all the conclusions concerning the value $I\{0\}$ of the functional which have been drawn above are automatically extended to functionals of form (12).

Now let us consider a functional similar to (7) but dependent on an arbitrary number n of unknown functions:

$$I\{\mathbf{y}\} = \int_a^b [\mathbf{y}'^* \mathbf{P}(x)\,\mathbf{y}' + \mathbf{y}^* \mathbf{Q}(x)\,\mathbf{y}]\,dx \tag{14}$$

Here $\mathbf{y} = \mathbf{y}(x)$ is a column vector composed of n elements which depend on x, $\mathbf{P}(x)$ and $\mathbf{Q}(x)$ are symmetric square matrices depending continuously on x, and the asterisk designates the operation of transposing a matrix (a column or row vector is understood, respectively, as a matrix of the type $(n \times 1)$ or $(1 \times n)$).

If the boundary conditions are of the form

$$\mathbf{y}(a) = 0, \quad \mathbf{y}(b) = 0$$

the more general functional containing an additional term of the form $\mathbf{y}'^* \mathbf{R}(x)\,\mathbf{y}$ under the integral sign is reduced to form (14) if the matrix $\mathbf{R}(x)$ is symmetric (how?). The system of Euler's equations for functional (14) can be written in the vector form

$$(\mathbf{P}(x)\,\mathbf{y}')' - \mathbf{Q}(x)\,\mathbf{y} = \mathbf{0} \tag{15}$$

(Let the reader prove the validity of vector equation (15) by passing to the scalar form of the functional and using the equality $\frac{d}{dy_i}\sum_{i,j} a_{ij} y_i y_j = 2\sum_j a_{ij} y_j$.)

Let the quadratic form $\boldsymbol{\xi}^* \mathbf{P}(x)\,\boldsymbol{\xi}$ be positive definite for every value of x under consideration. Denote by $\mathbf{Y}^k(x)$ $(k = 1, 2, \ldots, n)$ the solution of system (15) satisfying the initial conditions $\mathbf{y}(a) = 0$, $\mathbf{y}'(a) = \mathbf{C}^k$, where $\mathbf{C}^k$ $(k = 1, 2, \ldots, n)$ are arbitrary constant vectors such that the determinant of the matrix $(\mathbf{C}^1, \mathbf{C}^2, \ldots, \mathbf{C}^n)$ (whose columns are the vectors $\mathbf{C}^1, \mathbf{C}^2, \ldots, \mathbf{C}^n$) is different from zero. The first point $\tilde{a}$, lying to the right of the point $x = a$, at which $\det(\mathbf{Y}^1(x), \mathbf{Y}^2(x), \ldots, \mathbf{Y}^n(x)) = 0$ is said to be *conjugate* to a. (It can be proved that $\tilde{a}$ is independent of the particular choice of the vectors $\mathbf{C}^k$.) Reasoning as in the scalar case we can prove for the value $I\{\mathbf{0}\} = 0$ of the functional the same assertions as in the third paragraph of the present section.

5. Jacobi Conditions. The results established in Sec. 4 can be applied to expression (3) which is a quadratic functional with respect to δy. Let $\bar{y}(x)$ be a function which satisfies Euler's equation (1.36) and boundary conditions (1.26), i.e. let functional (1.18) take on a stationary value for $y = \bar{y}(x)$. For brevity, we put $\bar{F}''_{yy} = F''_{yy}(x, \bar{y}(x), \bar{y}'(x))$, etc. Then, by virtue of Sec. 3,

if the expression $\overline{F}''_{y'y'}$ changes its sign on the interval $a \leqslant x \leqslant b$, $I\{\bar{y}\}$ is a minimax value of the functional. Besides, as was mentioned in Sec. 1.15, the investigation of the case when $\overline{F}''_{y'y'}$ has zeros is connected with essential difficulties. Therefore, we shall suppose that $\overline{F}_{y'y'}$ retains its sign; for definiteness, let $\overline{F}''_{y'y'} > 0$ $(a \leqslant x \leqslant b)$.

Denoting temporarily δy by z in functional (3) (into which the function $\bar{y}(x)$ is substituted for y) we can write Euler's equation corresponding to the functional in the form

$$(\overline{F}''_{y'y'}z')' + \left(\frac{d}{dx}\overline{F}''_{yy'} - \overline{F}''_{yy}\right)z = 0 \qquad (16)$$

Relation (16) considered as a homogeneous linear differential equation of the second order with the unknown function $z(x)$ is called **Jacobi's equation** (for the original functional (1.18)). Let $\tilde{a}$ denote the point conjugate to a (for this equation). Note that, in contrast to the original quadratic functional, in the case of the general functional (1.18) this conjugate point depends not only on the form of the functional but also on the choice of the extremal $\bar{y}(x)$. If $\tilde{a} > b$, we see that, by virtue of Sec. 4, functional (3) has a minimum for δy 0. Consequently, according to Sec. 2, *the condition $\tilde{a} > b$ is sufficient for the functional $I\{y\}$ to have a weak minimum for $y = \bar{y}(x)$; if $\tilde{a} < b$, the functional $I\{y\}$ has a weak minimax for $y = \bar{y}(x)$.* **It** is evident that in the case $\overline{F}''_{y'y'} < 0$ $(a \leqslant x \leqslant b)$ we must speak of a maximum instead of a minimum. These sufficient conditions for extremum expressed in terms of conjugate points are known as the **Jacobi conditions.**

The Jacobi conditions admit of a simple geometrical interpretation. Consider the one-parameter family of extremals satisfying the first boundary condition (1.26), i.e. consisting of all extremals passing through the given point $A(a, y_A)$. A family of that kind is referred to as a **central field of extremals.** (In Fig. 96 we see a typical disposition of extremals forming a central field.) Fixing one of these extremals $\bar{y}(x)$, substituting the expression $y = \bar{y}(x) + \delta y$ into Euler's equation (1.37) and discarding the terms of order of smallness higher than the first relative to δy we just arrive at equation (16). Consequently, relation (16) is the linearized Euler's equation for extremals lying close to $\bar{y}(x)$. In particular, we see that $\tilde{a}$ is the abscissa of the first point of intersection, lying to the right of the point A, of the extremal $\bar{y}(x)$ with an extremal (belonging to the central field) which is infinitely close to $\bar{y}(x)$ (Fig. 96). Thus, $x = \tilde{a}$ is the abscissa of the first point (see

the point C in Fig. 96) of tangency of extremal $y=\bar{y}(x)$ to the envelope of the family of extremals. If $\overline{F''_{y'y'}}>0$ and the terminal point B_1 of the extremal $\bar{y}(x)$ lies to the left of C, functional I achieves a weak minimum on the extremal AB_1 (in comparison with the curves having the same end points and lying close to the extremal). But if the terminal point of the extremal occupies a position of the type of B_2, i.e. lies to the right of C, we can continuously pass to a piecewise smooth extremal of the type of

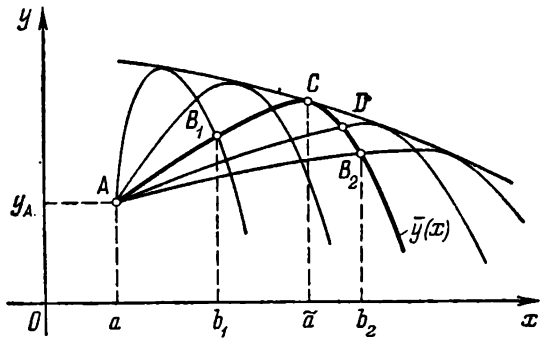


Fig. 96

ADB_2, etc. until we obtain the extremal AB_2, and, in this process, the value of the functional is decreased, which means that we have a minimax.

As an example, let us take the minimum-surface-of-revolution problem considered in Secs. 1.1 and 1.8. Here we have $F=2\pi y\sqrt{1+y'^2}$, that is, $F''_{y'y'}>0$ for $y>0$, and therefore the functional has no maximum values. Choosing a concrete extremal $y=k\cosh\frac{x}{k}$ (see formula (1.42)) we obtain (check it up!) Jacobi's equation

$$\left(\frac{k}{\cosh^2\frac{x}{k}}z'\right)'+\frac{1}{k}\frac{z}{\cosh^2\frac{x}{k}}=0 \qquad (17)$$

The function $z=\sinh\frac{x}{k}$ is a particular solution of this homogeneous linear equation, and therefore the order of the equation can be reduced by means of the substitution $z=u\sinh\frac{x}{k}$ (e.g. see [107], Sec. XV.14, Property 4). This results in the general solution

$$z=C_1\sinh\frac{x}{k}+C_2\left(\cosh\frac{x}{k}-\frac{x}{k}\sinh\frac{x}{k}\right)$$

of equation (17) (let the reader carry out all the calculations). If $a=-b<0$, the initial conditions $z(a)=0$, $z'(a)\neq 0$ are satisfied for $C_1=\dfrac{\cosh\frac{b}{k}}{\sinh\frac{b}{k}}-\dfrac{b}{k}$ and $C_2=1$, whence, after some transformations, we obtain

$$z=\frac{\sinh\frac{x+b}{k}}{\sinh\frac{b}{k}}-\frac{x+b}{k}\sinh\frac{x}{k}=\left(\frac{\sinh\frac{x+b}{k}}{\sinh\frac{b}{k}\sinh\frac{x}{k}}-\frac{x+b}{k}\right)\sinh\frac{x}{k}$$

Thus, the point $(-b)$ (conjugate to the point $-\tilde{b}$) is the first zero of the function $z(x)$ lying to the right of the point $x=-b$. The function is positive on the interval $-b<x\leqslant 0$, and the differentiation clearly indicates that the above expression in the brackets is a monotone function of x on the interval $0<x<\infty$ decreasing from ∞ to $-\infty$. Consequently, the inequalities $(-\tilde{b})>b$ and $(-\tilde{b})<b$ are, respectively, equivalent to the inequalities

$$\frac{\sinh\frac{2b}{k}}{\sinh\frac{b}{k}}-\frac{2b}{k}\sinh\frac{b}{k}>0 \quad \left(\text{that is} \quad \cosh\frac{b}{k}-\frac{b}{k}\sinh\frac{b}{k}>0\right)$$

and $\cosh\frac{b}{k}-\frac{b}{k}\sinh\frac{b}{k}<0$. In their turn, the latter inequalities are, respectively, equivalent to the requirement that the point $\left(b,\ k\cosh\frac{b}{k}\right)$ lie on the catenary $y=k\cosh\frac{x}{k}$ to the left of the point of tangency of this curve to the envelope of the family of extremals (see Fig. 93) and to the right of it. Consequently, one of the two catenaries passing through the point M (the point of intersection of the corresponding wire ring with the xy-plane; see Fig. 97), namely, curve I, generates the catenoid with minimum area, while the rotation of the other catenary (curve II) results in a minimax area.

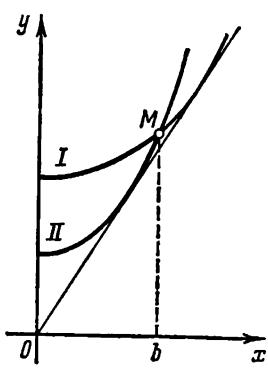


Fig. 97

It can be shown (but we do not present the proof here) that the former surface has the minimum area not only in comparison with other surfaces of revolution but also among all the surfaces (which are not necessarily axially symmetric) satisfying the given boundary conditions. This justifies the hypothesis that in the case

of axially symmetric wire rings the form of equilibrium of the soap film is also axially symmetric (see Sec. 1.1). This fact is by far not evident because, generally speaking, a problem involving some symmetry conditions may possess a nonsymmetric solution.

The Jacobi conditions for the functional

$$I\{\mathbf{y}\} = \int_a^b F(x, \mathbf{y}, \mathbf{y}')\,dx, \quad \mathbf{y}(a) = \mathbf{y}_a, \quad \mathbf{y}(b) = \mathbf{y}_b \tag{18}$$

depending on a vector function $\mathbf{y}(x)$, that is on several scalar functions, are derived in the same manner. In this case we obtain, instead of (3), the expression

$$\delta^2 I\{\mathbf{y}, \delta\mathbf{y}\} = \int_a^b \left[\delta\mathbf{y}^* \left(F''_{y_i y_j}\right) \delta\mathbf{y} + 2\delta\mathbf{y}^* \left(F''_{y_i y'_j}\right) \delta\mathbf{y}' + \right.$$
$$\left. + \delta\mathbf{y}'^* \left(F''_{y'_i y'_j}\right) \delta\mathbf{y}'\right] dx \tag{19}$$

for the second variation. If the matrix $\left(F''_{y_i y'_j}\right)$ is symmetric, the sign of quadratic functional (19) can be determined with the aid of the results presented at the end of Sec. 4.

6. Geodesics. Here we shall discuss a consequence of the Jacobi conditions. Suppose that $F''_{y'y'} > 0$ $(a \leqslant x \leqslant b)$ along an extremal $y = \bar{y}(x)$ and that the coefficients of equation (16) are bounded. Then *every sufficiently small portion of the extremal* $\bar{y}(x)$ *corresponding to an interval* $a_1 \leqslant x \leqslant b_1$ $(a \leqslant a_1,\ b_1 \leqslant b)$ *minimizes the functional* $\int_{a_1}^{b_1} F(x, y, y')\,dx$ *(considered under the boundary conditions* $y(a_1) = \bar{y}(a_1)$, $y(b_1) = \bar{y}(b_1)$*)*. The same assertion applies to functionals dependent on several functions, for instance, of the type of (1.46), if all eigenvalues of the matrix $\left(F''_{y'_i y'_j}\right)$ are positive at every point of the extremal.

To prove the latter assertion we estimate every term of the quadratic form $2\delta\mathbf{y}^* \left(F''_{y'_i y'_j}\right) \delta\mathbf{y}'$ under the integral sign in (19) according to the formula

$$\mathbf{A}\delta y_i \delta y'_j \geqslant -\frac{|A|}{2\varepsilon} (\delta y_i)^2 - |\mathbf{A}| \frac{\varepsilon}{2} (\delta y'_j)^2 \tag{20}$$

where $\varepsilon > 0$ is an arbitrary positive number (let the reader prove inequality (20)). Then we combine the sum of the first terms appearing on the right-hand side of formula (20) with the first quadratic form under the integral sign in (19), and the sum of the second terms with the third quadratic form in (19). If we fix a sufficiently small positive value of ε, the resultant third

quadratic form remains positive definite because the property of a quadratic form to be positive definite is stable with respect to small variations of its coefficients (why?). Consequently, by virtue of the assertions concerning functional (14) for sufficiently small values of the difference $b-a$ (see Sec. 4), the functional obtained after inequality (20) has been applied, as described above, takes on only positive values. Hence, the original functional possesses the same property.

For a functional $\int\limits_{(G)} F(\mathbf{x}, y, y'_{\mathbf{x}})\,dG$ dependent on a function of several variables we can formulate a similar proposition. In this case, for the functional to have a minimum it is necessary that the eigenvalues of the matrix $(F''_{p_i p_j})$ be nonnegative, and it is sufficient that these eigenvalues be positive and the domain (G) sufficiently small.

These propositions establish sufficient conditions guaranteeing the existence of "local" extrema attained when intervals (or domains) of integration are sufficiently small. They admit of a visual geometric interpretation connected with the notion of **a geodesic (geodesic line)** on a given surface (S), that is a line whose arc length between its two arbitrary points has a stationary value. In other words, if we give a sufficiently small variation to an arbitrary arc (l_{AB}) of such a line without changing the position of the end points A and B and without falling outside the surface (S), the increment of the arc length is of higher order of smallness relative to the variation of the arc. (The variation of the arc is measured in the sense of C_1, that is when we say that the variation is small we mean that the points of the arc (l_{AB}) receive small displacements and that the change of the direction of the tangent line at each point of the arc is also small.) An arbitrary variation of a finite arc AB can be obtained as a superposition of variations of small portions of that arc (in Fig. 98 these small portions are AB_1, A_1B_2 and A_2B), and therefore the above definition is equivalent to the requirement that the arc lengths of small portions of a geodesic line assume stationary values. (By the way, this argument is of general character and can be applied to other variational problems.) But, given a small portion of an arc lying on a surface (S), we can always choose Cartesian coordinates such that the equation of the corresponding part of the surface is written in the form $z=\varphi(x, y)$ and the equations of the arc (lying on that part of the surface) in the form $y=y(x)$, $z=\varphi(x, y(x))$ (why?). Consequently, we can limit oursel-

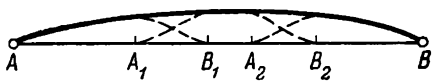


Fig. 98

ves to discussing a stationary value of the functional

$$\int_a^b \sqrt{dx^2+dy^2+dz^2} = \int_a^b \sqrt{1+y'^2+[\varphi'_x(x,\ y)+\varphi'_y(x,\ y)\,y']^2}\,dx \quad (21)$$

considered under given initial conditions $y(a)=y_a$ and $y(b)=y_b$. Euler's equation (1.36) corresponding to (21) is a second-order differential equation, and therefore there is exactly one geodesic line passing through every point of the surface (S) in any given direction. Here we have $F''_{y'y'}>0$ (check it up!), and consequently, on every sufficiently small arc of a geodesic line the arc length not only assumes a stationary value but also its minimum value (in Sec. 7 we shall even prove that this minimum is strong).

At the same time, the arc length of a finite (not small) portion of a geodesic line does not necessarily have the minimum value and may assume a minimax value. This fact can be demonstrated by an example in which the role of the surface (S) is played by a sphere of radius R whose geodesics are the arcs of great circles obtained in the sections of the sphere by the planes passing through its centre. If an arc of that kind is smaller than the semicircular arc of radius R, its length assumes the minimum value in comparison with the lengths of all the other curves lying on the sphere sufficiently close to that arc and having the same end points. But if this arc is greater than the semicircle of radius R, its length assumes a minimax value because we can both increase and decrease the length by means of an appropriate infinitesimal deformation without changing the position of its end points.

Let now (S) be an arbitrary surface and (l) be a geodesic starting from a point A of (S). To determine the maximal portion of the geodesic line on which the minimum of the arc length is attained we can apply the Jacobi conditions which enable us to draw the conclusion that the terminal point of this maximal portion is the first point $\tilde{A}$ of intersection of the geodesic (l) with another geodesic line starting from A in a direction which is infinitely close to the direction of (l) at the point A. This point $\tilde{A}$ (conjugate to A) separates the arcs of (l) starting at A and having minimum length from the arcs with minimax lengths. If the point A does not exist, the arc length has minima on all the arcs of the geodesic line (l) which start from the point A.

Geodesic lines demonstrate another peculiarity of many variational problems. Let us consider two points A and B on the surface of a right circular cylinder (Fig. 99). When an arbitrary surface is subject to bending (without changing the arc lengths on it) the geodesic lines go into the geodesic lines (why?). Therefore, if the cylinder is developed, rolled out, on the plane the geode-

sics of the cylinder go into the geodesics of the plane, the latter being straight lines. Consequently, the geodesic lines of the right circular cylinder are circular helices with any lead h and their limiting forms which are circles (with $h=0$) and straight lines ($h=\infty$). But any two points A and B on the surface of the cylinder not lying in one section perpendicular to the axis of the cylinder can be connected by an infinitude of arcs of circular helices (see Fig. 99 where three such helices are depicted), the length of each arc assuming a minimum value (in comparison with all the other curves with the same end points lying sufficiently close to the arc under consideration). Thus, the problem of minimum for the length of the arc joining the two points has infinitely many solutions among which there is one corresponding to the absolute minimum (this is arc I in Fig. 99) while all the other correspond to local minima. There is a wide class of variational problems which, like the above problem, do not possess a unique solution, although this is by far not always so (for example, this is not the case in the problems enumerated in Sec. 1).

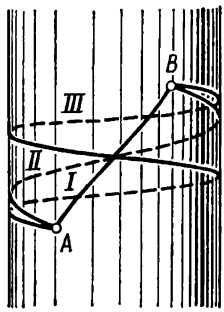


Fig. 99

There is an interesting mechanical interpretation of a geodesic line on a surface (S). Let us write functional (21) in the parametric form

$$\int \sqrt{\dot{x}^2+\dot{y}^2+\dot{z}^2}\,dt$$

The equation of the surface (S) which we take in the form $H(x, y, z)=0$ serves as a finite constraint here. According to Sec. 1.13, in order to obtain the geodesics, we must form the system of Euler's equations corresponding to the auxiliary function

$$F^*=\sqrt{\dot{x}^2+\dot{y}^2+\dot{z}^2}-\lambda(t)\,H(x, y, z)$$

This system can be written in the scalar form

$$\left.\begin{aligned}
-\lambda(t)\,H'_x-\frac{d}{dt}\frac{\dot{x}}{\sqrt{\dot{x}^2+\dot{y}^2+\dot{z}^2}}&=0\\
-\lambda(t)\,H'_y-\frac{d}{dt}\frac{\dot{y}}{\sqrt{\dot{x}^2+\dot{y}^2+\dot{z}^2}}&=0\\
-\lambda(t)\,H'_z-\frac{d}{dt}\frac{\dot{z}}{\sqrt{\dot{x}^2+\dot{y}^2+\dot{z}^2}}&=0
\end{aligned}\right\}$$

or in the vector form

$$\frac{d\tau}{dt}+\lambda(t)\ \text{grad}\ H=0$$

(where τ is the unit tangent vector to the sought-for line (L)) which is obtained if we multiply the above equations by the corresponding unit vector of the coordinate axes and add the results together. As is known (e.g. see [107], Secs. VII.24 and XII.2), the vector $\frac{d\tau}{dt}$ goes along the principal normal of (L), and the vector grad$H$ is in the direction of the normal to (S). Thus, we conclude that the principal normal of a geodesic line on a surface is normal to the surface at each point. But if a point moves without tangential acceleration along a trajectory its acceleration is directed along the principal normal of the trajectory, that is the acceleration only has the centripetal (normal) component. Consequently, when a point is in "inertial motion" in a surface, in the sense that there are no external forces and friction, its trajectory is a geodesic line.

The notion of a geodesic line can be similarly introduced for manifolds of Euclidean spaces of any dimension and also for general Riemannian spaces (see Secs. V.2.4, 5). These geodesics whose properties are the same as in the case of a two-dimensional surface play an important role in the general theory of manifolds and spaces analogous to the role of straight lines in Euclidean spaces. This question will be discussed again in § 4.

7. Conditions for Strong Extremum. In this section we shall derive some sufficient conditions for strong extremum (see Sec. 1.3) of functional (1.18) with boundary condition (1.26). Let $\overline{y}(x)$ $(a\leqslant x\leqslant b)$ be an extremal satisfying the boundary conditions and the Jacobi conditions described in Sec. 5 which are sufficient for weak extremum. These conditions are "almost necessary" in the sense that there is only one ambiguous case arising when the strict inequalities in the Jacobi conditions may turn into equalities. On the other hand, as was shown in Sec. 1.3, there are cases in which a weak minimum is not a strong one, and therefore it becomes clear that the Jacobi conditions are not sufficient for strong minimum. But there is an additional condition (established by K. Weierstrass) which, together with the Jacobi conditions, guarantees the existence of a strong minimum (and is "almost necessary" in the above sense).

Let us introduce the so-called **Weierstrass function**

$$E(x, y, p, q)=F(x, y, q)-F(x, y, p)-(q-p)F'_p(x, y, p)$$

(If we put $q=p+h$, it is readily seen that this expression is

the nonlinear part of the expansion of the function $F(x, y, p+h)$ in powers of h.) The additional condition mentioned above is written in the form

$$E(x, \bar{y}(x), \bar{y}'(x), q) > 0 \qquad (a \leqslant x \leqslant b) \tag{22}$$

and should be fulfilled for all $q \neq \bar{y}'(x)$ (it is obvious that $E(x, y, p, p) \equiv 0$). Furthermore, if

$$E(x, \bar{y}(x), \bar{y}'(x), q) < 0 \tag{23}$$

for at least one combination of values of x and q there is no strong minimum for $y = \bar{y}$.

In particular, condition (22) holds if $F''_{pp}(x, y, p) > 0$ for $y = \bar{y}(x)$ and for all the values of p, which is implied by the formula

$$\varphi(q) - \varphi(p) - (q-p)\varphi'(p) = \int_p^q dr \int_p^r \varphi''(s)\, ds$$

where φ is an arbitrary function (let the reader derive this formula). By the way, it follows that the assertion (mentioned in Sec. 6) that on every sufficiently small portion of a geodesic line the arc length reaches a strong minimum is true.

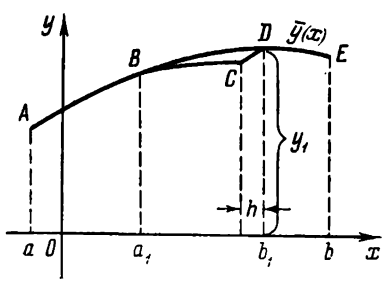


Fig. 100

The proof of these propositions concerning the Weierstrass function can be found in courses of calculus of variations. Here we shall limit ourselves to some general considerations elucidating this question. Suppose that an arc BD of the extremal $\bar{y}(x)$ (see Fig. 100) is replaced by a curve BCD with corner point C, the slope of the part CD being equal to q and sufficiently differing from the slope $p = \bar{y}'(b_1)$ of the extremal at the point D. Then the curve $ABCDE$ is close to the extremal AE relative to the metric of $C[a, b]$ but not in the sense of $C_1[a, b]$. Putting, for brevity, $x_D - x_C = h$ and applying formula (1.91) to the interval $a_1 \leqslant x \leqslant b_1$ we receive the expression of δI corresponding to this variation of the extremal:

$$\delta I = F'_{y'}(b_1, y_1, p)(-qh) + [F(b_1, y_1, p) - pF'_{y'}(b_1, y_1, p)](-h) + F(b_1, y_1, q)h = E(b_1, \bar{y}(b_1), \bar{y}'(b_1), q)h$$

Hence, if condition (23) holds for certain x and q, we can take this value of x as b_1 and thus decrease the value $I\{\bar{y}\}$. This

means that there is no strong minimum for $y=\bar{y}$. But if condition (22) is fulfilled for all x and q, any deformation (of the extremal) of the above type leads to an increase of $I\{\bar{y}\}$. Combining variations of this form we can obtain any line belonging to a sufficiently wide class of curves which are close to the original extremal not only in the sense of C but also relative to the metric of C_1. Finally, to variations which are small in the sense of C_1, we can apply the Jacobi conditions, which makes it possible to complete the proof.

In the case of functional (18) we can similarly formulate the sufficient conditions for strong minimum by using the Weierstrass function of the form

$$E\{x, \mathbf{y}, \mathbf{p}, \mathbf{q}\} = F\{x, \mathbf{y}, \mathbf{q}\} - F\{x, \mathbf{y}, \mathbf{p}\} - (\mathbf{q}^* - \mathbf{p}^*) F'_{\mathbf{p}}(x, \mathbf{y}, \mathbf{p})$$

In conclusion we note that a maximum of $I\{y\}$ is a minimum of $-I\{y\}$, and therefore, in what follows, we can limit ourselves to investigating minima of functionals.

8. Variational Theory of Eigenvalues. Quadratic functional (7) with a function $P(x)$ satisfying condition (9) can be investigated with the aid of another approach which is extremely useful for some problems and completely analogous to the method of studying quadratic forms and self-adjoint mappings on the basis of extremal properties of their eigenvalues (see Secs. IV.1.6, 7). Here we shall consider functional (7) for boundary conditions (8).

Based on the above-mentioned analogy with the theory of quadratic forms, we shall consider the values of the functional (7) for the functions $y=y(x)$ satisfying the integral constraint

$$\int_a^b y^2 dx = 1 \tag{24}$$

These values of (7) are bounded below because, under conditions (9) and (24), we have the inequality

$$I\{y\} \geqslant \int_a^b Q(x) y^2 dx \geqslant \min_{a \leqslant x \leqslant b} Q(x) \int_a^b y^2 dx = \min_{a \leqslant x \leqslant b} Q(x)$$

Let $y_1(x)$ be the function satisfying (24) for which $I\{y\}$ attains the minimum value equal to $\min_{a \leqslant x \leqslant b} Q(x)$. (The fact that such a function exists needs a proof because, generally speaking, a functional whose values are bounded below does not necessarily attain the minimum value, and, to improve the situation, we sometimes have to introduce generalized solutions; see Sec. 1.20. Some details concerning this question will be discussed in Sec. 9.) According to the rule given in Sec. 1.12, the function $y_1(x)$

satisfies the equation

$$(P(x)y_1')' - [Q(x)y_1 - \lambda_1 y_1] = 0 \tag{25}$$

Putting, for brevity,

$$L[y] = -(P(x)y')' + Q(x)y \tag{26}$$

we can rewrite equation (25) in the form

$$L[y_1] = \lambda_1 y_1 \tag{27}$$

If the expression $L[y]$ is interpreted as a linear operator in the corresponding functional space, formula (27) indicates that y_1 is an eigenvector (in this case also termed an **eigenfunction** of this operator) belonging to the eigenvalue λ_1.

For our further aims we need the formula

$$\int_a^b (Lu)v\,dx = \int_a^b uLv\,dx \tag{28}$$

which is valid for any two functions $u(x)$ and $v(x)$ satisfying boundary conditions (8) (let the reader derive this formula with the aid of integration by parts).

Strictly speaking, here we must only take the functions possessing the derivatives entering into expression (26) which must be finite or such that all the integrals involved are absolutely convergent. In what follows we shall not make stipulations of this kind. It is convenient to interpret these functions as elements of Hilbert's space $L_2[a, b]$ (e.g. see [107], Sec. XVII.28). Formula (28) then means that the operator L considered on the subspace of functions satisfying conditions (8) is self-adjoint (see Sec. IV.1.6).

After the function $y_1(x)$ has been constructed we proceed to consider the minimum of functional (7) under the same boundary conditions and two integral constraints of the form

$$\int_a^b y^2\,dx = 1, \quad \int_a^b y_1(x)\,y\,dx = 0 \tag{29}$$

According to Sec. 1.11, the function $y_2(x)$ for which this minimum is reached must satisfy the equation $L[y_2] = \lambda_2 y_2 + \mu_2 y_1$ (check it up!). Multiplying both members of this equation by y_1, integrating with respect to x from a to b and using formulas (27), (28) and (29) we conclude that $\mu_2 = 0$ (let the reader carry out the computations). Hence, $y_2(x)$ is an eigenfunction of the operator L corresponding to the eigenvalue λ_2.

On the third eigenfunction $y_3(x)$ associated with the eigenvalue λ_3 functional (7) achieves a minimum under boundary condi-

tions (8) and the integral constraints

$$\int_a^b y^2\,dx=1, \quad \int_a^b y_1(x)\,y\,dx=0 \quad \text{and} \quad \int_a^b y_2(x)\,y\,dx=0$$

etc. Proceeding in this way we thus obtain an infinite sequence

$$y_1(x),\ y_2(x),\ y_3(x),\ \ldots \tag{30}$$

of eigenfunctions of the operator L which satisfy boundary conditions (8). By virtue of the constraints imposed on these functions, they are all mutually orthogonal and satisfy condition (24) (which is a **normalization condition**; the functions satisfying this condition are said to be **normalized**).

We obviously have $I\{y_1\}\leqslant I\{y_2\}\leqslant I\{y_3\}\leqslant\ldots$ because when we impose an additional constraint the class of functions under consideration becomes narrower, and therefore the least value of the functional attained on this class can only increase. (Let the reader carefully examine this argument which has a general significance.) It can be shown (but here we do not present the proof) that the sequence $I\{y_n\}$ is infinitely large, that is tends to infinity. Finally, performing integration by parts in the integral of the expression $P(x)y'^2$ (see formula (7)) we can readily deduce the formula

$$I\{y\}=\int_a^b L[y]\,y\,dx+P(x)\,yy'\Big|_a^b \tag{31}$$

from which, on the basis of equalities $L[y_n]=\lambda_n y_n$ $(n=1, 2, \ldots)$ and relation (24), we derive the relations

$$I\{y_n\}=\int_a^b \lambda_n y_n y_n\,dx=\lambda_n \quad (n=1,\ 2,\ \ldots) \tag{32}$$

The set of the eigenvalues λ_n is called the **spectrum** of the operator L (considered under conditions (8)).

Sequence (30) possesses another important property: it is not only orthogonal but also *complete* in the sense that any function (belonging to a wide class of functions) can be expanded into series with respect to this system. The proof of the completeness can be found in courses devoted to integral equations and spectral theory of differential operators. Here we shall only give some explanatory considerations concerning this question. Like in Sec. 1.12, we can represent every function $y(x)$ by a set y_0, y_1, ..., y_n of its values (or, which is more convenient, by the set $\sqrt{h}y_k$, $k=1, 2, \ldots, n$) where h is the interval of calculations (step) in the x direction). Then functional (7) goes into

a quadratic form of these values, and the procedure described above goes into the procedure of determining the eigenvectors of the matrix of a quadratic form (see Sec. IV.1.7), the set of these eigenvectors being complete (in the corresponding finite-dimensional space).

For example, let us consider the functional

$$I\{y\} = \int_0^l y'^2\, dx \tag{33}$$

under boundary conditions (8) of the form

$$y(0) = y(l) = 0 \tag{34}$$

Here the equation for the eigenfunctions takes the form $-y'' = \lambda y$, $y(0) = y(l) = 0$. Solving this equation (let the reader check up the result!) we arrive at the following sequence of eigenfunctions and eigenvalues:

$$y_k(x) = \sqrt{\frac{2}{l}} \sin \frac{k\pi}{l} x, \quad \lambda_k = \left(\frac{k\pi}{l}\right)^2 \qquad (k = 1, 2, 3, \ldots) \tag{35}$$

(the factor $\sqrt{\frac{2}{l}}$ is the **normalization coefficient** introduced in order to make the norms of the functions be equal to unity). A series with respect to these eigenfunctions is nothing but Fourier's sine series (e.g. see [107], formula (XVII.107)).

It should be stressed that in the foregoing discussion of completeness the approximating quadratic form must be considered not in the whole $(n+1)$-dimensional space of $(n+1)$-tuples $(y_0, y_1, \ldots, y_n)$ but in its $(n-1)$-dimensional subspace determined by the equalities $y_0 = 0$ and $y_n = 0$. Therefore, when expanding a function with respect to the eigenfunctions, we most often suppose that it satisfies conditions (8), that is belongs to the corresponding subspace of the functional space in question. For instance, as is known (e.g. see [107], Sec. XVII.25), for Fourier's sine series (i.e. a series expansion with respect to the eigenfunctions found in the above example) to be uniformly convergent we not only require that the expanded function $f(x)$ should be continuous but also that the equalities $f(0) = f(l) = 0$ be fulfilled. For general functional (7) and boundary conditions (8) the situation remains the same. (By the way, to guarantee the uniform convergence of the series obtained by termwise differentiation of the expansion of the function $f(x)$ we must additionally require that $f(x)$ should be continuously differentiable, etc.) On the other hand, we can also consider a more general type of convergence, namely, convergence in the mean square, that is suppose that the expansion of the function $f(x)$ converges relative to the metric of the space

$L_2[a, b]$. Then, as in the case of Fourier's series, conditions (8) can be violated and the function $f(x)$ can even be discontinuous; the only essential condition that should be imposed in this case is that $f(x)$ is a square-summable function.

The operator L and its eigenvalues can be considered for boundary conditions different from (8). In this case the variational theory is also applicable. For instance, if we do not impose a boundary condition for the function on which functional (7) depends at an end point of the interval in question (or at both end points), we can realize the above construction if we appropriately take into account that, by virtue of Sec. 1.15, the eigenfunctions must satisfy natural boundary condition (1.82) at that end point (for functional (7) the natural boundary condition simply reduces to the relation y' 0). All the above considerations and properties remain valid for this case as well.

If we take the more general functional

$$I\{y\} = \int_a^b [P(x)y'^2 + Q(x)y^2]\,dx + \alpha[y(a)]^2 + \beta[y(b)]^2 \qquad (36)$$

(instead of (7)) and consider it without boundary conditions, the results of Sec. 1.13 imply that the natural boundary conditions (see (1.84)) are of the form

$$(P(x)y' - \alpha y)_{x=a} = 0, \quad (P(x)y' + \beta y)_{x=b} = 0 \qquad (37)$$

The scheme of constructing eigenvalues and eigenfunctions remains unchanged for this case (but, of course, we must minimize functional (36), instead of (7)). It should be noted that for functions satisfying conditions (37) formula (28) as well as the resultant formula (32) in which I is computed according to formula (36) remain valid (let the reader check up these properties with the aid of integration by parts).

At first glance it seems that it is possible to obtain the same result if we minimize functional (7) under condition (24) for the class of functions satisfying boundary conditions (37), then introduce the orthogonality condition and so on. But this conclusion is not true. The matter is that every function $y = y(x)$ (say, belonging to C_1) can be given an infinitesimal variation (resulting in an infinitesimal variation of functional (7)) such that the new function satisfies conditions (37). Indeed, given the end point values of the function y, we can compute the end point values of y' determined by formula (37) and then appropriately change the values of $y(x)$ in the vicinity of the end points in such a way that the deviation of the new graph from the old one is small and the slopes of the tangent lines to the new graph at the end points are equal to the values of y' thus found (see

Fig. 101). Therefore, in principle, there is no difference between determining the least values of functional (7) for the class of functions satisfying conditions (37) and for all the functions belonging to the functional space we deal with. On the other hand, the function $y_0(x)$ minimizing functional (7) considered on the class of all functions subject to constraint (24) satisfies conditions $y'(a)=y'(b)=0$. Consequently, functional (7) considered on the class of functions satisfying conditions (37) does not attain the least value. If we take an appropriately chosen sequence of functions satisfying (37) for which the values of the functional tend to its greatest lower bound on that class of functions (e.g. cf. [107], Sec. IV.19) we obtain, in the limit, the function $y_0(x)$ which does not satisfy conditions (37). (Let the reader find out why this argument is inapplicable to the case of boundary conditions (8).)

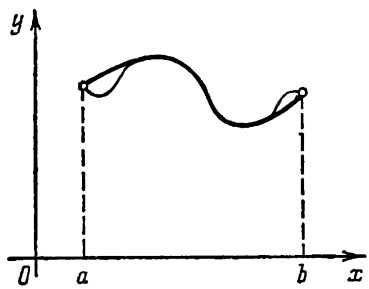


Fig. 101

The variational theory of eigenvalues is similarly constructed for operators applied to functions of several independent variables. For instance, the minimization of the quadratic functional

$$I\{z\}=\int\limits_{(G)}\left[\left(\frac{\partial z}{\partial x}\right)^2+\left(\frac{\partial z}{\partial y}\right)^2+Q(x, y)\,z^2\right]dG \tag{38}$$

under the constraint

$$\int\limits_{(G)} z^2\,dG=1 \tag{39}$$

and the boundary condition

$$z|_{(\Gamma)}=0 \tag{40}$$

leads to the first eigenfunction $z_1(x, y)$ and the least eigenvalue λ_1 of the operator

$$L[z]=-\frac{\partial^2 z}{\partial x^2}-\frac{\partial^2 z}{\partial y^2}+Q(x, y)\,z \tag{41}$$

considered on the class of functions satisfying condition (40). The additional constraint $\int\limits_{(G)} zz_1\,dG=0$ results in the second eigenfunction belonging to the subsequent eigenvalue, etc. (it is meant that the eigenvalues are arranged into an increasing sequence although in problems involving functions of two arguments it is more natural to supply the eigenvalues and the eigenfunctions

with two indices). It can be shown that all the basic assertions formulated above for functional (7) remain true for this case as well.

If on a part of the contour (Γ) (or on the whole contour) there is no boundary condition, the eigenfunctions satisfy, on that part, the natural boundary condition (see (1.101))

$$\frac{\partial z}{\partial x}\cos(\widehat{\mathbf{n}, x})+\frac{\partial z}{\partial y}\cos(\widehat{\mathbf{n}, y})=0, \text{ that is } \frac{\partial z}{\partial n}=0$$

Analogously, if we consider the functional

$$\int\limits_{(G)}\left[\left(\frac{\partial z}{\partial x}\right)^2+\left(\frac{\partial z}{\partial y}\right)^2+Q(x, y)z^2\right]dG-\int\limits_{(\Gamma)}\varphi(x, y)z^2\,d\Gamma$$

(instead of (38)) without boundary conditions, the application of this scheme leads to eigenfunctions of operator (41) which satisfy the natural boundary conditions of the form

$$\left[\frac{\partial z}{\partial n}+\varphi(x, y)z\right]_{(\Gamma)}=0$$

9. Existence of Minimum. Let us dwell in more detail on the question of existence of a function minimizing a given functional. The same question arises in many other similar problems.

For definiteness, consider functional (7):

$$I\{y\}=\int\limits_a^b[P(x)y'^2+Q(x)y^2]\,dx$$

We shall suppose that $P(x)>0$ and take the boundary conditions (of type (8))

$$y(a)=0, \quad y(b)=0$$

and the integral constraint (24):

$$\int\limits_a^b y^2\,dx=1$$

We have already mentioned that the values of the functional are bounded below. Therefore, even if its least value is not attained, we can construct a sequence of functions $Y_n(x)$ satisfying conditions (8) and (24) for which $\lim\limits_{n\to\infty} I\{Y_n\}$ is equal to the greatest lower bound of the values of the functional. A sequence of this type is referred to as a **minimizing sequence** for the functional I. (We have an analogous situation for the number sequence 1, $\frac{1}{2}$, ..., $\frac{1}{n}$, ...: the set of all positive numbers does not contain the least number but the limit of the above sequence,

consisting of positive numbers, is equal to zero; the sequence $\left\{\frac{\sqrt{2}}{n^2}\right\}$ also possesses this property, etc. If the set of all positive numbers coincides with the set of all the values of a functional, these sequences can be regarded as sequences of values of the functional assumed on functions forming minimizing sequences.)

If a minimizing sequence $\{Y_n(x)\}$ (or its subsequence) is convergent (say, uniformly convergent) to a limit function $Y(x)$, the functional achieves the sought-for least value for $y=Y(x)$. Indeed, we have

$$\int_a^b Y^2(x)\,dx = \lim_{n\to\infty} \int_a^b Y_n^2(x)\,dx = 1$$

and

$$\int_a^b Q(x)\,Y^2(x)\,dx = \lim_{n\to\infty} \int_a^b Q(x)\,Y_n^2(x)\,dx$$

while

$$\int_a^b P(x)\,Y'^2(x)\,dx \leqslant \lim \int_a^b P(x)\,Y_n'^2\,dx$$

(The latter relation seems strange because, at first glance, one may think that the right member must be equal to the left

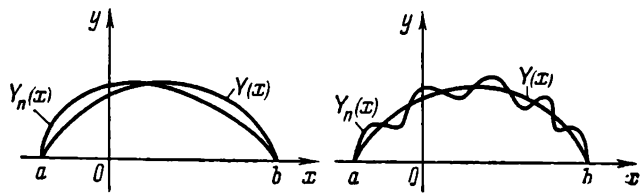


Fig. 102

member. But there are in fact cases when the right member exceeds the left one; both possibilities are demonstrated in Fig. 102.) Consequently, we have

$$I\{Y\} \leqslant \lim_{n\to\infty} I\{Y_n\}$$

The sequence $\{Y_n(x)\}$ being minimizing, we thus conclude that, under the given conditions, the quantity $I\{Y\}$ is the least value of the functional. (Let the reader carefully examine this argument.)

Conditions for the existence of a convergent sequence of the type of $\{Y_n\}$ involve the notion of *compactness* which plays a very important role in these and many other problems. *A set (M) of*

points (elements) of a normed space (M) (or of a more general metric space (R)) is said to be (sequentially) compact in (R) if every infinite sequence of points belonging to (M) contains at least one convergent subsequence. Using this notion we can restate the Bolzano-Weierstrass theorem (see page 89) very briefly as follows: Any bounded set of points belonging to a finite-dimensional Euclidean space is compact. On the other hand, it is also true that every unbounded subset of a finite-dimensional Euclidean space is noncompact. To prove this assertion, P_0 is assumed to be chosen, as in the text, to be the origin of coordinates. Then by a finite-dimensional Euclidean space we mean the totality of all (M) which are assumed to be bounded sets.

A [illegible] of infinite-dimensional spaces is that these bounded subsets are not necessarily compact and therefore the investigation of these spaces becomes more complicated for instance, the sequence of functions $\sin nx$ [illegible]. A function has been considered as a point of the space $C[a, b]$ for any fixed a and b (for the reader familiar with this concept we should like to remark that the noncompactness of the sequence is due to the fact that the values of the functions $\sin x, \sin 2x, \sin 3x, \ldots$ oscillate between the finite limits -1 and $+1$ with increasing frequency, as $x \to \infty$, and therefore the subsequences of this sequence are not uniformly convergent. But at the same time it is possible to prove that every infinite sequence $\{f_n(x)\}$ [illegible] above and can be bounded by the same constant in the space $C[a, b]$ because in this case the absolute values of all the [illegible] remain smaller than unity, that is, $|f_n(x)| < 1$ for all n and all x in $[a, b]$. Moreover, the sequence $\{f_n(x)\}$ [illegible] the inequality

$$\left| \int_a^x f_n(x)\,dx \right| < \text{[illegible]}$$

[illegible] and therefore the sequence $\{f_n(x)\}$ could be extracted a [illegible] These properties guarantee the existence of a uniformly convergent subsequence of the sequence $\{f_n(x)\}$ whose limit function is to be shown above to be the one for which the functional attains the least value.

Of course, this is only an alternate argument, and to prove the existence of [illegible] it is necessary to carry out a more careful analysis (cf. [107]). But nevertheless the general scheme presented above includes most of the basic ideas characteristic of such a more rigorous investigation and can also be applied to many other problems related to problems of variational problems. These indications should be sufficient to enable the reader to understand a [illegible] and prove the completeness of the sequence in an appropriate (Hilbert) space. Although these considerations often fail to provide a practical method by means of which at least approximately the [illegible] they contain

the correctness of the formulation of the problem and impel us to seek effective approximate and numerical methods for determining the solution and investigating its properties.

10. Basic Condition for Minimum. The basic condition for existence of a minimum of a quadratic functional can be formulated in terms of the properties of eigenvalues, which yields some propositions of general importance.

Here we shall suppose that one of the coefficients entering into each boundary condition (37) is equal to zero. Then, by virtue of (31), we can write the relation

$$I\{y\} = \int_a^b L[y]\, y\, dy \tag{42}$$

for every function $y(x)$ satisfying these conditions.

Taking advantage of the fact that, under conditions (37), system (30) of eigenfunctions of the operator L is complete, we can write down the expansion of the function $y(x)$ into series with respect to this system:

$$y(x) = \xi_1 y_1(x) + \xi_2 y_2(x) + \ldots = \sum_{n=1}^{\infty} \xi_n y_n(x) \tag{43}$$

The coefficients ξ_n entering into (43) are found by means of the ordinary formulas

$$\xi_n = \int_a^b y y_n\, dx, \quad n = 1, 2, \ldots$$

for a series in orthogonal functions (e.g. see [107], formulas (XVII.100); in this particular case the denominators $\int_a^b y_n^2\, dx$, $n = 1, 2, \ldots$, are equal to unity, according to the normalization condition for the eigenfunctions). Substituting expansion (43) into (42), performing multiplication and using the orthogonality and normalization conditions for the eigenfunctions we obtain the formula

$$I\{y\} = \int_a^b \sum_{n=1}^{\infty} \xi_n L[y_n] \sum_{m=1}^{\infty} \xi_m y_m\, dx =$$

$$= \sum_{n,m=1}^{\infty} \xi_n \xi_m \int_a^b \lambda_n y_n y_m\, dx = \sum_{n=1}^{\infty} \lambda_n \xi_n^2 \tag{44}$$

The function $y(x)$ (satisfying the boundary conditions) being taken quite arbitrarily, the coefficients ξ_n in (44) can also be regarded as independent arbitrary quantities. This enables us to conclude that *if all the eigenvalues* λ_n *are positive, functional* (42)

achieves the absolute minimum, under conditions (37), for the function $y = 0$; *if there is at least one negative eigenvalue the functional has a minimax for* $y = 0$. We have $\lambda_n \xrightarrow[n\to\infty]{} \infty$, and therefore the least eigenvalue λ_{min} is sure to exist (in the case under consideration we simply have $\lambda_{min} = \lambda_1$). This makes it possible to express the above conclusions in terms of the least eigenvalue: if $\lambda_{min} > 0$ we have the absolute minimum of the functional, and if $\lambda_{min} < 0$ we have a minimax. (It is also evident that if $\lambda_{min} = 0$, the functional has a nonstrict minimum for $y = 0$.)

As was mentioned, the above proof is of a general character, and therefore the conclusions we have drawn also apply to many other classes of quadratic functionals; this becomes particularly clear if, at the very beginning, we write the functional in the form of the right-hand side of (42). The most important condition for these conclusions to be valid is that the operator of the type of L is self-adjoint under the chosen boundary conditions, that is formula (28) must hold (this condition guarantees that the eigenvalues are real). Another important condition is that the corresponding system of eigenfunctions is complete. In particular, under certain assumptions, these conclusions are valid for functionals of functions dependent on several variables to which the Jacobi conditions are no longer applicable.

For functionals I of a more general type (which are not necessarily quadratic) the above conclusions can be applied, by analogy with Sec. 5, to the second variation $\delta^2 I$, which makes it possible to express sufficient (and also necessary) conditions for weak local minimum in terms of the eigenvalues of the operator corresponding to linearized Euler's equation. (The only essential distinction of this case from the foregoing one is that the condition $\lambda_{min} > 0$ does not necessarily imply that there is a nonstrict minimum because here we must take into account the contributions made by the terms of higher order of smallness.)

A real extremum problem for a functional usually involves certain parameters on which the eigenvalues of the corresponding operator depend continuously. Of course, in the general case this functional relationship may have discontinuities, but, as a rule, when there is a discontinuity corresponding to a combination of values of the parameters, the variational problem itself degenerates in a certain sense. The simplest case arises here when there is only one parameter α. Then, after the dependence of λ_{min} on α is determined (usually, by means of numerical methods), we find the corresponding zeros and points of discontinuity (provided there are such), which enables us to specify the possible critical values of α for which λ_{min} changes the sign. These changes of sign obviously correspond to passages from minimum values to

minimax values and vice versa. If there are two parameters we can similarly determine the corresponding lines in the plane of the parameters on which these changes occur. The situation becomes more complicated when the number of parameters exceeds two, although even in that case we can try to derive an approximate formula for hyperplanes on which there are zeros and discontinuities by means of interpolation formulas.

Relation (44) is analogous to the formula of reduction of a quadratic form to its principal axes (see Sec. IV.2.1): it reduces the quadratic functional $I\{y\}$ to diagonal form after we have passed, in the space $L_2[a, b]$, to a Euclidean basis composed of the corresponding eigenfunctions of form (30). But, in contrast to the finite-dimensional case, this reduction does not necessarily apply to the whole space and may only hold for the subspace of the functions y for which the series on the right-hand side of formula (44) is convergent. This is accounted for by the fact that, as is implied by (32), the functional $I\{y\}$ is unbounded in the general case (see Sec. 1).

For instance, let us consider functional (33) under boundary conditions (34). The eigenvalues and eigenfunctions of the functional are given by formula (35). Here, to guarantee the convergence of series (44), we require that the function (35) be continuous and satisfy the same boundary conditions (34) because, if otherwise, as is known (e.g. see [107], Sec. XVII.25), the coefficient ξ_n is of the order of $\frac{1}{n}$, and therefore the product $\lambda_n \xi_n^2$ does not tend to zero for $n \to \infty$. Besides, for the values of the functional $I\{y\}$ to be finite we must assume that the function $y(x)$ possesses a square-summable derivative. It is the collection of all functions $y(x)$ satisfying these conditions that forms the subspace for which formula (44) is valid.

It may seem, at first glance, that the above example involves a contradiction: the left-hand member of (44) is equal to zero for $y \equiv 1$ whereas the right-hand member turns into infinity. The explanation of this paradox lies in the fact that when expansion (43) is applied under boundary conditions (34) we must take, instead of $y \equiv 1$, the function

$$y(x) = \begin{cases} 1 \text{ for } 0 < x < \pi \\ 0 \text{ for } x = 0 \\ 0 \text{ for } x = \pi \end{cases}$$

whose derivative y' contains terms with delta function. These terms are not square-summable, and therefore both members of equality (44) simultaneously turn into infinity.

In the minimum problems for the more general quadratic functionals considered in this section the subspaces for which the reduction

to principal axes is valid are also composed of functions with square-summable derivatives, the boundary conditions being specified by the formulations of the original problems.

11. Dependence of Eigenvalues on the Functional. Let us take functional (7) with boundary conditions (8). We have shown that the corresponding operator of type (26) has, under these conditions, the sequence of eigenfunctions of form (30) and the corresponding sequence of eigenvalues of form (32). If the coefficients $P(x)$ and $Q(x)$ or the limits of integration a, b or the boundary conditions are changed the eigenvalues gain some variations. We can pose the problem of determining these variations. Here we shall discuss some cases when it is possible to answer this question.

If the coefficients P and Q receive nonnegative variations, that is the new coefficients $\tilde{P}(x)$ and $\tilde{Q}(x)$ satisfy the inequalities

$$\tilde{P}(x) \geqslant P(x), \quad \tilde{Q}(x) \geqslant Q(x) \qquad (a \leqslant x \leqslant b) \tag{45}$$

every new eigenvalue $\tilde{\lambda}_n$ $(n = 1, 2, \ldots)$ exceeds the corresponding eigenvalue λ_n provided that relations (45) *do not simultaneously turn into equalities.*

The method of constructing the first eigenvalue given in Sec. 8 immediately implies that the assertion is true for λ_1:

$$\tilde{\lambda}_1 = \tilde{I}\{\tilde{y}_1\} > I\{\tilde{y}_1\} \geqslant I\{y_1\} = \lambda_1$$

(Let the reader carefully examine this chain of inequalities!)

To prove the assertion for the nth eigenvalue we can apply the following theorem (**Courant's theorem** for functionals): *let $\rho_1(x), \ldots, \rho_{n-1}(x)$ be an arbitrary set of $n-1$ functions* and $\lambda(\rho_1, \rho_2, \ldots, \rho_{n-1})$ *be the minimum value of functional* (7) *attained on the class of all functions satisfying conditions* (8) *and* (24) *and orthogonal to each function $\rho_i(x)$ $(i = 1, 2, \ldots, n-1)$; then the greatest of the numbers $\lambda(\rho_1, \rho_2, \ldots, \rho_{n-1})$, considered for all the possible combinations of the functions $\rho_1(x), \ldots, \rho_{n-1}(x)$, is equal to the eigenvalue λ_n* (this proposition generalizes Courant's theorem for quadratic forms, given in Sec. IV.1.7, to quadratic functionals).

If a constant is added to the function $Q(x)$ all the eigenvalues receive the increment equal to this constant. In fact, the function $y(x)$ being subject to constraint (24) introduced in Sec. 8, we have

$$\int_a^b [P(x)y'^2 + (Q(x) + C)y^2]\,dx = \int_a^b [P(x)y'^2 + Q(x)y^2]\,dx + C$$

and thus the constant C is added to all the values of functional (7), which implies the validity of the last assertion.

These properties enable us to derive a crude estimate for the eigenvalues, namely

$$\min P(x)\left(\frac{n\pi}{b-a}\right)^2+\min Q(x)\leqslant\lambda_n\leqslant$$

$$\leqslant\max P(x)\left(\frac{n\pi}{b-a}\right)^2+\max Q(x)\qquad(n=1,\ 2,\ 3,\ \ldots)\quad(46)$$

Indeed, if, for instance, we replace, by virtue of the first property, the coefficients $P(x)$ and $Q(x)$ by their greatest values and then apply the second property to the term involving y^2 and take advantage of formulas (35) established for functional (33), this results in the upper bound for λ_n indicated in formula (46) (let the reader carefully analyse this argument!).

To derive a more precise estimation of the eigenvalues we make a change of variables

$$y=\varphi(x)\,z,\quad dx=\psi(x)\,dt$$

such that equation (24) of the constraint and the orthogonality conditions of type (29) retain their form while functional (7) goes into an integral of the form

$$\int_0^l\left[z'^2+r(t)\,zz'+s(t)\,z^2\right]dt=\int_0^l\left[z'^2+\left(s(t)-\frac{1}{2}r'(t)\right)z^2\right]dt$$

The first requirement implies that $\varphi^2\psi=1$, and the second that $\frac{P\varphi^2}{\psi}=1$ (check it up!). It follows that

$$\varphi(x)=[P(x)]^{-\frac{1}{4}},\quad\psi(x)=\sqrt{P(x)}$$

that is

$$t=\int_a^x[P(x)]^{-\frac{1}{2}}dx,\quad l=\int_a^b[P(x)]^{-\frac{1}{2}}dx$$

and, besides, we have $s-\frac{1}{2}r'=\frac{1}{4}P''-\frac{1}{16}P^{-1}P'^2+Q$ (check up the calculations!). Now, noting that the eigenvalues remain invariant under this change of variables we can apply formula (46) and the above assertions concerning variations of the eigenvalues, which results in

$$\left(\frac{n\pi}{\int_a^b P^{-\frac{1}{2}}dx}\right)^2+\min\left(\frac{1}{4}P''-\frac{1}{16}P^{-1}P'^2+Q\right)\leqslant\lambda_n\leqslant$$

$$\leqslant\left(\frac{n\pi}{\int_a^b P^{-\frac{1}{2}}dx}\right)^2+\max\left(\frac{1}{4}P''-\frac{1}{16}P^{-1}P'^2+Q\right)\qquad(n=1,\ 2,\ 3,\ \ldots)$$

The distinction between this estimate and inequality (46) lies in the fact that here the difference between the rightmost and leftmost members is bounded for $n \to \infty$.

If a is increased or b is decreased all the eigenvalues become greater.

Indeed, for definiteness, suppose that $\tilde{a} = a$, and $\tilde{b} > b$. Let us extend the function $y_1(x)$ to the whole interval $\tilde{a} \leqslant x \leqslant \tilde{b}$ by putting $y_1 \equiv 0$ for $b \leqslant x \leqslant \tilde{b}$. Then, denoting the extended function by $z(x)$, we can write

$$\lambda_1 = I(y_1) = \tilde{I}(z) > \tilde{I}(\tilde{y}_1) = \tilde{\lambda}_1$$

For the subsequent eigenvalues the corresponding estimations are obtained in like manner with the aid of Courant's theorem.

Now suppose that we do not introduce a boundary condition on one or both end points of the interval $[a, b]$. Then, as was mentioned, the eigenfunctions must satisfy the condition $y' = 0$ at that end point. It is readily shown that all the above assertions concerning eigenvalues are valid for these boundary conditions as well, and the proof also remains almost the same. In addition to the properties of eigenvalues stated above we can formulate the following assertion: *if, on one of the end points a and b, the condition $y = 0$ is replaced by the condition $y' = 0$, all the eigenvalues are decreased.* Actually, according to the foregoing assertion, this replacement is equivalent to discarding the boundary condition for the corresponding end point; then we obviously pass to a wider class of functions for which the values of functional (7) are compared, and therefore the least value of the functional is decreased.

Analogous properties of eigenvalues hold for some other problems in which the variational methods can be used for constructing eigenvalues. For instance, applying similar considerations to functional (38) we can establish the following result: *if the domain (G) is replaced by its subdomain, all the eigenvalues of operator* (41) *considered with boundary conditions* (40) *are increased.*

§ 3. CANONICAL EQUATIONS AND VARIATIONAL PRINCIPLES

1. Canonical Form of Euler's Equations. In many theoretical studies it is convenient to transform Euler's equations to the *canonical form*, which makes it possible to elaborate a general approach to various classes of variational problems. This proves particularly useful in some problems of physics and mechanics where the new variables introduced for canonical transformations can be given an obvious physical interpretation.

We shall consider stationary and, in particular, extremal values of the functional

$$I\{y_1, y_2, \ldots, y_n\} = \int_a^b F(x, y_1, y_2, \ldots, y_n, y_1', y_2', \ldots, y_n')\,dx \tag{1}$$

where the functions $y_1(x), \ldots, y_n(x)$ satisfy certain boundary conditions. According to Sec. 1.10, for this functional the system of Euler's equations is written in the form

$$F'_{y_i} - \frac{d}{dx}F'_{y_i'} = 0 \qquad (i = 1, 2, \ldots, n) \tag{2}$$

and is a system of n ordinary differential equations of the second order with n unknown functions

$$y_1(x), y_2(x), \ldots, y_n(x) \tag{3}$$

Let us introduce the quantities

$$p_i = F'_{y_i'}(x, y_1, y_2, \ldots, y_n, y_1', y_2', \ldots, y_n') \quad (i = 1, 2, \ldots, n) \tag{4}$$

which, together with the variables y_i $(i = 1, 2, \ldots, n)$ are called **canonical (Hamiltonian) variables** (corresponding to functional (1)). The variables y_i and p_i are said to be **canonically conjugate.** For any arbitrarily chosen functions (3), variables (4) are also certain functions of x.

Consider the function

$$H(x, y_1, \ldots, y_n, p_1, \ldots, p_n) = -F(x, y_1, \ldots, y_n, y_1', \ldots, y_n') + \sum_{i=1}^{n} y_i' p_i \tag{5}$$

where the derivatives y_i' entering into the right-hand side are regarded as being expressed in terms of $x, y_1, \ldots, y_n, p_1, \ldots, p_n$ by means of relations (4) (which can be considered to be a system of n finite equations in n unknowns $y_1', \ldots, y_n'$). The function $H(x, y_1, \ldots, y_n, p_1, \ldots, p_n)$ is called the **Hamiltonian function** associated with functional (1).

As is known (e.g. see [107], Secs. IX.13 and XII.3), the theory of implicit functions states that if the corresponding Jacobian of a system of finite equations is distinct from zero the solution (at least local) of the system exists. In the case under consideration this Jacobian is $\det\left(F''_{y_i'y_j'}\right)$; as was indicated in Sec. 1.19, the condition that this determinant is different from zero plays an essential role in variational problems.

Let the reader show that the Hamiltonian function corresponding to functional (1.102) has the form

$$H = -\sqrt{1-p^2}$$

Differentiating relation (5) and using equalities (4) we readily obtain the formula

$$dH = -F'_x dx - \sum F'_{y_i} dy_i - \sum F'_{y'_i} d(y'_i) + \sum d(y'_i)\, p_i + \\ + \sum y'_i dp_i = -F'_x dx - \sum F'_{y_i} dy_i + \sum y'_i dp_i$$

Thus, if the variable x and the canonical variables y_i, p_i $(i = 1, 2, \ldots, n)$ are (functionally) independent, we must have

$$\frac{\partial H}{\partial x} = -F'_x, \quad \frac{\partial H}{\partial y_i} = -F'_{y_i}, \quad \frac{\partial H}{\partial p_i} = y'_i \tag{6}$$

Now we see that if functions (3) specify an extremal of functional (1), i.e. satisfy Euler's equations (2), the second and the third equalities (6) imply that

$$\frac{dp_i}{dx} = -\frac{\partial H}{\partial y_i}, \quad \frac{dy_i}{dx} = \frac{\partial H}{\partial p_i} \qquad (i = 1, 2, \ldots, n) \tag{7}$$

System (7) is referred to as the **canonical form of Euler's equations** or the **canonical (Hamiltonian) system of differential equations** corresponding to functional (1). It is a system of $2n$ ordinary differential equations of the first order with $2n$ unknown functions $y_i(x)$, $p_i(x)$ $(i = 1, 2, \ldots, n)$.

Note that, according to the Weierstrass-Erdmann conditions, p_i $(i = 1, \ldots, n)$ and H are the quantities which must vary continuously together with x and y_i $(i = 1, \ldots, n)$ when the variable point passes through a corner point of a piecewise smooth extremal. This becomes particularly clear if, by analogy with (1.91), we rewrite the expression for an arbitrary variation of functional (1) in the vector form

$$\delta \int_a^b F(x, \mathbf{y}, \mathbf{y}')\, dx = \int_a^b \left(F'_{\mathbf{y}} - \frac{d}{dx} F'_{\mathbf{y}'}\right) \cdot \delta \mathbf{y}\, dx + (\mathbf{p} \cdot \delta[\mathbf{y}(x)] - H\, \delta x)\Big|_{x=a}^{b} \tag{8}$$

and take into account the method of deriving the Weierstrass-Erdmann conditions used in Sec. 1.20.

2. First Integrals. Let $\Phi(x, y_1, y_2, \ldots, y_n, p_1, p_2, \ldots, p_n)$ be an arbitrary function. If we consider its values along a solution of the system of equations (7), that is substitute into it the solution of the system for the variables y_i, p_i $(i = 1, \ldots, n)$, we

obtain a composite function of the variable x for which

$$\frac{d\Phi}{dx}=\frac{\partial\Phi}{\partial x}+\sum_i\frac{\partial\Phi}{\partial y_i}\frac{dy_i}{dx}+\sum_i\frac{\partial\Phi}{\partial p_i}\frac{dp_i}{dx}=$$

$$=\frac{\partial\Phi}{\partial x}+\sum_{i=1}^{n}\left(\frac{\partial\Phi}{\partial y_i}\frac{\partial H}{\partial p_i}-\frac{\partial\Phi}{\partial p_i}\frac{\partial H}{\partial y_i}\right) \tag{9}$$

The expression

$$\sum_{i=1}^{n}\left(\frac{\partial\Phi}{\partial y_i}\frac{\partial H}{\partial p_i}-\frac{\partial\Phi}{\partial p_i}\frac{\partial H}{\partial y_i}\right)$$

is termed the **Poisson bracket of the functions Φ and H** and is denoted by the symbol $[\Phi, H]$. We thus can write formula (9) as

$$\frac{d\Phi}{dx}=\frac{\partial\Phi}{\partial x}+[\Phi, H] \tag{10}$$

Let the reader prove the following simple properties of the Poisson brackets:

$$[\Phi, \Psi]=-[\Psi, \Phi],\ [\Phi, \Phi]\equiv 0,\ [\Phi_1+\Phi_2, \Psi]=[\Phi_1, \Psi]+[\Phi_2, \Psi],$$
$$[C\Phi, \Psi]=C[\Phi, \Psi]\ (C=\text{const}),$$
$$[\Phi_1\Phi_2, \Psi]=\Phi_1[\Phi_2, \Psi]+\Phi_2[\Phi_1, \Psi]$$

In particular, it follows from (10) that for the relation

$$\Phi(x, y_1, \ldots, y_n, p_1, \ldots, p_n)=\text{const}$$

to be fulfilled along every solution of system (7) it is necessary and sufficient that the identity

$$\frac{\partial\Phi}{\partial x}+[\Phi, H]\equiv 0 \tag{11}$$

hold. A relation of the form $\Phi(x, y_1, \ldots, y_n, p_1, \ldots, p_n)=\text{const}$ is called a **first integral** of system (7); as is known (e.g. see [107], Sec. XV.13), if we are given k independent first integrals, it is possible to reduce by k the number of the equations in the system.

Now let us turn to the particular case when the function H does not explicitly involve the variable x, that is has the form $H=H(y_1, \ldots, y_n, p_1, \ldots, p_n)$. Then if we put $\Phi=H$, relation (11) is obviously fulfilled. Consequently, if the function H is independent of x it is constant along any integral curve of system (7), that is assumes a constant value for all the points of any extremal (but, of course, it can take on different values on different extremals). (Let the reader prove this fact directly by computing the total derivative $\frac{dH}{dx}$ along an arbitrary integral

curve of equations (7).) If $n=1$ we just arrive at first integral (1.38).

Another simple case, when a first integral exists, appears if the function H does not depend on a variable y_i or p_i; we then can, respectively, put $\Phi=p_i$ or $\Phi=y_i$ (check it up!).

3. Canonical Transformations. Canonical equations can be considered irrespective of the original variational problem for which the quantities p_i have been introduced. We can consider these equations in the $(2n+1)$-dimensional space x, y_i, p_i $(i=1, \ldots, n)$; if H is independent of x we can also regard them as determining trajectories in the $2n$-dimensional space of the variables y, p $(y=(y_1, \ldots, y_n), p=(p_1, \ldots, p_n))$ with x as parameter. In other words, we simply deal with a system of differential equations of form (7) in which H is a given function of $2n+1$ variables x, y, p which determines the right members of equations (7). This approach appears particularly natural if we note that the variables y and p are involved symmetrically in system (7) and, in this sense, their roles are quite equivalent.

Of course, by far not every system of ordinary differential equations in the space x, y, p has a canonical form. If we are given an arbitrary system of differential equations

$$\frac{dy_i}{dx}=\Phi_i(x, y_1, \ldots, y_n, p_1, \ldots, p_n), \quad \frac{dp_i}{dx}=\Psi_i(\ldots) \qquad (i=1, \ldots, n)$$

it is canonical if and only if

$$\frac{\partial\Phi_i}{\partial p_j}\equiv\frac{\partial\Phi_j}{\partial p_i}, \quad \frac{\partial\Psi_i}{\partial y_j}\equiv\frac{\partial\Psi_j}{\partial y_i}; \quad \frac{\partial\Phi_i}{\partial y_j}+\frac{\partial\Psi_j}{\partial p_i}\equiv 0 \qquad (i, j=1, 2, \ldots, n)$$

(why?). Therefore, if we make a change of variables

$$\begin{aligned} Y_i&=g_i(x, y_1, \ldots, y_n, p_1, \ldots, p_n),\\ P_i&=h_i(x, y_1, \ldots, y_n, p_1 \ldots, p_n) \qquad (i=1, \ldots, n) \end{aligned} \tag{12}$$

in canonical system (7), the system of differential equations in the new variables x, Y, P $(Y=(Y_1, \ldots, Y_n), P=(P_1, \ldots, P_n))$ thus obtained is not necessarily canonical. But if the transformed system retains its canonical form, formulas (12) specify a **canonical transformation** (more precisely, a canonical transformation for original system (7) because if we apply the same formulas (12) to another canonical system its canonicity can be violated; a canonical transformation is sometimes referred to as a *contact transformation*, the latter term being elucidated in Sec. 4).

Here we shall describe an important class of canonical transformations. For this purpose we shall take advantage of the fact that system (7) can be interpreted as the system of Euler's

equations corresponding to the functional

$$I\{y,\ p\}=\int_a^b\left(\sum_i p_i y_i'-H\right)dx \tag{13}$$

(check it up!). As was mentioned in Sec. 1.7, every change of variables in a system of Euler's equations yields the same result as if we introduced the new variables into the functional and then formed Euler's equations for the transformed functional. This implies that if, after a change of variables of form (12) has been performed, functional (13) takes the form

$$\int_a^b\left(\sum_i P_i Y_i'-\tilde{H}(x,\ Y,\ P)\right)dx+\text{const}$$

(here we again denote by Y and P the collections of the variables $(Y_1,\ \ldots,\ Y_n)$ and $(P_1,\ \ldots,\ P_n)$), then transformation (12) is canonical. This is obviously the case if transformation (12) satisfies the identity

$$\sum p_i\,dy_i-H(x,\ y,\ p)\,dx=\sum P_i\,dY_i-\tilde{H}(x,\ Y,\ P)\,dx+d\Phi(x,\ y,\ Y) \tag{14}$$

where $\Phi(x,\ y,\ Y)$ is a function called the **generating function** of canonical transformation (12). (Note that we have $\int d\Phi=\Phi\Big|_a^b$, and, since the end point values of the sought-for functions determining the extremals can be regarded as being given, the last term in formula (14) only results in a constant summand in the functional, which is inessential.) Given the generating function, we can construct all the transformations belonging to the class we are interested in. For, if we consider x, y and Y as being independent variables, relation (14) shows that

$$p_i=\frac{\partial\Phi}{\partial y_i},\quad P_i=-\frac{\partial\Phi}{\partial Y_i}\quad(i=1,\ 2,\ \ldots,\ n)\quad \tilde{H}=H+\frac{\partial\Phi}{\partial x} \tag{15}$$

and, conversely, formulas (15) imply (14). Hence, to obtain a canonical transformation we can choose a function $\Phi(x,\ y,\ Y)$ (which, in the general case, can be taken quite arbitrarily), express Y in terms of x, y, p by means of the first n equations (15) (provided this is possible), substitute these expressions into the subsequent n equations (15) and thus construct a canonical transformation of form (12).

It is sometimes more convenient to take a generating function Φ dependent on the variables x, y, P. Then, if instead of (15) we take the formulas

$$p_i=\frac{\partial\Phi}{\partial y_i},\quad Y_i=\frac{\partial\Phi}{\partial P_i}\qquad(i=1,\ 2,\ \ldots,\ n),\quad \tilde{H}=H+\frac{\partial\Phi}{\partial x}$$

we arrive at the identity

$$\sum p_i dy_i - H\,dx = \sum P_i dY_i - \tilde{H}\,dx + d\left(\Phi - \sum P_i \frac{\partial \Phi}{\partial P_i}\right)$$

which replaces (14) in this case. (Let the reader check up this result.) The latter identity shows that the transformation $y,\ p \longrightarrow Y,\ P$ thus obtained is also canonical.

The theory of canonical transformations becomes simpler when the Hamiltonian function H does not depend on x. In this case it appears natural to assume that the functions $g_i,\ h_i\ (i = 1, \ldots, n)$ in formulas (12) and also the function Φ in (14) are independent of x as well. Then relations (15) indicate that $\tilde{H} = H$, that is, under these canonical transformations, the Hamiltonian function remains invariant. Besides, it follows from formula (14) that the difference

$$\sum p_i dy_i - \sum P_i dY_i$$

coincides with the total differential of the function $\Phi(y, Y)$. It can be shown (but here we do not present the proof) that for the equality $d(\sum p_i dy_i - \sum P_i dY_i) = d\Phi$ to hold it is necessary and sufficient that the relations

$$[Y_i, Y_j] = 0,\ [P_i, P_j] = 0,\ [P_i, Y_j] = 0 \qquad (i \neq j),\ [P_i, Y_i] = 1$$

be fulfilled (the square brackets denote the Poisson brackets introduced in Sec. 2). Finally, if the last relations take place, the Poisson bracket $[\Psi_1, \Psi_2]$ taken for two arbitrary functions Ψ_1 and Ψ_2 assumes the same value when it is expressed in the variables $y,\ p$ and in $Y,\ P$, and thus is invariant with respect to these canonical transformations.

4. Contact Transformations. The first $2n$ formulas (15) can be given an interesting geometric interpretation. Consider two n-dimensional spaces y and Y (for simplicity's sake we shall denote the spaces and their elements by the same letters y and Y). The vectors $y = (y_1, \ldots, y_n)$ and $Y = (Y_1, \ldots, Y_n)$ of these spaces may depend on x but we shall consider x to be fixed and therefore this dependence will not be indicated. Let us associate, with each point (vector) $y = (y_1, y_2, \ldots, y_n)$ of the first space, an $(n-1)$-dimensional surface (hypersurface) lying in the second space which is described by an equation of the form

$$\Phi(y, Y) = C \tag{16}$$

where C is a fixed constant. Now suppose that the point y travels along an arbitrary infinitesimal path, the infinitesimal displacement vector dy being perpendicular to a fixed direction specified by unit vector α^0. Surface (16) then also moves in the space, and if dy takes on all the infinitesimal vector values perpendicular

to α^0, this results in the appearance of an $(n-1)$-parameter family of surfaces (because dy has $(n-1)$ degrees of freedom), with "infinitesimal width", intersecting at a point Y (there can be several points Y of that kind). The position of the point Y depends both on the choice of the point y and on the choice of the direction of unit vector α^0. Let $\mathbf{A}^0$ be a unit vector in the space Y drawn through that point Y of intersection of the surfaces perpendicularly to them. A nonzero vector lying in an n-dimensional space uniquely determines the $(n-1)$-dimensional hyperplane to which it is perpendicular. Therefore a pair (y, α^0) formed of the point y and the vector α^0 can be interpreted as a point with which an element of the corresponding $(n-1)$-dimensional hyperplane (perpendicular to α^0) is associated (the same, of course, applies to every pair $(Y, \mathbf{A}^0)$). We shall simply refer to (y, α^0) and $(Y, \mathbf{A}^0)$ as *elements*. Using this terminology we can say that the above construction specifies a transformation under which elements (y, α^0) go into elements $(Y, \mathbf{A}^0)$; this is not an ordinary mapping which transforms points into points because it is applied to objects of a more complicated nature (which we have called elements). A transformation of this type is termed a **contact transformation**, and the function Φ (or, more precisely, the function $\Phi - C$) is the **generating function** of the transformation.

Take a smooth $(n-1)$-dimensional hypersurface (s) in the space y; it possesses a tangent hyperplane at each (nonsingular) point, and thus the surface (s) generates the corresponding set of elements (y, α^0), every element being thought of as a point of the surface (s) with an element of the tangent hyperplane (determined by the corresponding normal unit vector α^0). Applying a contact transformation to these elements we obtain the corresponding set of elements $(Y, \mathbf{A}^0)$ in the space Y. Consequently, this contact transformation results in a point-to-point transformation of the surface (s) into the surface (S) consisting of the points Y entering into the above elements $(Y, \mathbf{A}^0)$ which are the images of (y, α^0). (It is clear that, under such a transformation, to a singular point of the surface (s), for instance, to a conic point, there corresponds a manifold of points lying on the surface (S).) If we apply the transformation to two surfaces (s_1) and (s_2) tangent to one another at a point y, the surfaces (S_1) and (S_2) (the images of (s_1) and (s_2)) are tangent at the point Y into which the point y goes under the transformation. But if the surfaces (s_1) and (s_2) intersect at a nonzero angle at y, there will be two different points $Y_1 \in (S_1)$ and $Y_2 \in (S_2)$ corresponding to the point y.

Let the reader investigate the contact transformation with generating function

$$\Phi = (Y_1 - y_1)^2 + (Y_2 - y_2)^2 + (Y_3 - y_3)^2$$

Here we have $n=3$, that is the spaces y and Y are three-dimensional, and it is convenient to think of them as coincident.

The construction of a contact transformation is closely related to the theory of envelopes (e.g. for the case $n=2$ see [107], Sec. XII.5). Let us consider a k-parameter family of $(n-1)$-dimensional surfaces $(1 \leqslant k \leqslant n-1)$ lying in a space of arbitrary dimension n, the equation of the family being of the form

$$F(Y_1 Y_2, \ldots, Y_n, C_1, C_2, \ldots, C_k) = \text{const} \qquad (17)$$

where $C_1, \ldots, C_k$ are the parameters. By analogy with the case $n=2$, it is readily shown that, generally speaking, the envelope

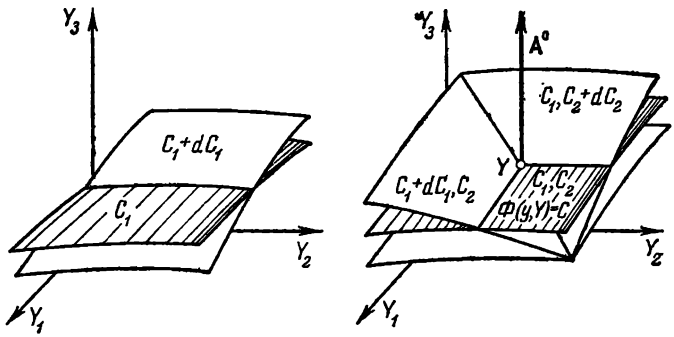


Fig. 103

of the family is (except some singular cases) an $(n-1)$-dimensional surface whose equation is obtained by eliminating the parameters from equation (17) (describing the family) and the equations

$$F'_{C_i}(Y_1, Y_2, \ldots, Y_n, C_1, C_2, \ldots, C_k) = 0 \qquad (i=1, \ldots, k) \quad (18)$$

Every surface belonging to family (17) touches the envelope along an $(n-1-k)$-dimensional manifold appearing in the intersection of that surface with the surfaces of the family lying infinitely close to it (by the way, it is sufficient to take the intersection with k surfaces each of which is obtained by means of an infinitesimal variation of one of the parameters $C_1, \ldots, C_k$; the cases $n=3$, $k=1$ and $n=3$, $k=2$ are demonstrated in Fig. 103 where the second figure simultaneously serves as an illustration of a contact transformation of an element for $n=3$). In particular, if $k=n-1$, then, generally speaking, every surface belonging to the family touches the enveloping surface at one or several discrete points (instead of adjoining it along a manifold).

Equations (18) can be written in the shortened form $\partial_C F = 0$ where $C=(C_1, C_2, \ldots, C_k)$ and $\partial_C F = F'_{C_1} dC_1 + F'_{C_2} dC_2 + \ldots + F'_{C_n} dC_n$. Therefore, coming back to relation (16) we see that the contact

transformation generated by the function $\Phi(y, Y)$ can be obtained if we consider equation (16) together with the vector condition

$$\partial_y \Phi(y, Y) = 0 \quad \text{for all} \quad dy \perp \alpha^0$$

or, equivalently, with the condition that

$$\operatorname{grad}_y \Phi(y, Y) \text{ is parallel to } \alpha^0 \tag{19}$$

where the symbol grad_y designates the operation of taking the gradient in the space y for a fixed element Y of the second space. The normal direction to surface (16) (the vector $\mathbf{A}^0$) is determined by the condition that

$$\mathbf{A}^0 \text{ is parallel to } \operatorname{grad}_Y \Phi(y, Y) \tag{20}$$

(e. g. see [107], Sec. XII.2). For any given y and α^0, relations (16) and (19) form a system of n scalar equations in n unknowns Y_1, $Y_2, \ldots, Y_n$. After these unknowns have been determined, we can find $\mathbf{A}^0$ from (20). It should be noted that, since the elements of both spaces y and Y are involved symmetrically in all the relations we have obtained, the contact transformation from the space y into the space Y we have considered simultaneously determines a contact transformation from Y into y with the same generating function (16).

Applying these results to canonical transformation (15) we conclude that its generating function Φ specifies a contact transformation from the space y into the space Y such that every element $(y, \mathbf{p}^0)$ goes into the corresponding element $(Y, \mathbf{P}^0)$.

5. Noether Theorem. In this section we consider a theorem which makes it possible to obtain first integrals of systems of Euler's equations for functionals possessing some invariance properties, which is important for many problems both in theoretical and practical aspects.

We begin with introducing the notion of a *continuous group of transformations* of the space E_n (or of a domain lying in this space) into itself. For brevity, we shall regard every point $(x_1, x_2, \ldots, x_n)$ of the space as being represented by its radius vector $\mathbf{x}$. Then transformation formulas

$$\bar{x}_i = \Phi_i(x_1, x_2, \ldots, x_n) \qquad (i = 1, 2, \ldots, n) \tag{21}$$

determining a mapping of E_n into itself can be written in the shortened form

$$\bar{\mathbf{x}} = \Phi(\mathbf{x}) \tag{22}$$

(One must not think that formula (22) says that rectilinear segments remain rectilinear under the mapping; this formula only expresses the law of transformation of the termini of the radius vectors expressed by equalities (21).)

Consider a one-parameter family of one-to-one mappings (transformations)

$$\bar{\mathbf{x}} = \mathbf{\Phi}(\mathbf{x}, t) \tag{23}$$

where t is a real scalar parameter. This family is called a **continuous group of transformations** if it possesses the following *group property*: the result of two consecutively applied transformations corresponding to any values t_1 and t_2 of the parameter is equivalent to the mapping, with parameter $t_1 + t_2$, belonging to the same family. In particular, it follows that the mapping with the value $t = 0$ of the parameter is the identity mapping, that is $\mathbf{\Phi}(\mathbf{x}, 0) \equiv \mathbf{x}$. It also follows that two mappings $\mathbf{\Phi}(\mathbf{x}, t)$ and $\mathbf{\Phi}(\mathbf{x}, -t)$ are mutually inverse (why?). For example, a mapping of the type $\bar{\mathbf{x}} = \mathbf{\Phi}(\mathbf{x}, t)$ may reduce to the translation of the space along a fixed direction by a distance of t or (e.g. for $n = 3$) to the rotation through an angle of magnitude t about a fixed axis; in the general case $\bar{\mathbf{x}} = \mathbf{\Phi}(\mathbf{x}, t)$ can be a nonlinear mapping. It is convenient to interpret the parameter t as time and family (23) of mappings as a flux in the space E_n. Then the group property simply means that the flux is stationary. The vector field

$$\mathbf{\Phi}_0(\mathbf{x}) = \mathbf{\Phi}'_t(\mathbf{x}, 0) \tag{24}$$

is the velocity field of the flux, and for every small Δt we have

$$\Delta\bar{\mathbf{x}} = \mathbf{\Phi}(\mathbf{x}, t + \Delta t) - \mathbf{\Phi}(\mathbf{x}, t) = \mathbf{\Phi}(\bar{\mathbf{x}}, \Delta t) - \bar{\mathbf{x}} = \mathbf{\Phi}_0(\bar{\mathbf{x}})\,\Delta t + o(\Delta t) \tag{25}$$

This accounts for the term the **infinitesimal generator** of group (23) applied to the auxiliary mapping $\bar{\mathbf{x}} = \mathbf{\Phi}_0(\mathbf{x})$ which is independent of t. Conversely, relation (25) implies that family (23) is obtained from its infinitesimal generator by means of solving the system of differential equations written in the vector form

$$\frac{d\bar{\mathbf{x}}}{dt} = \mathbf{\Phi}_0(\bar{\mathbf{x}}), \qquad \bar{\mathbf{x}}\Big|_{t=0} = \mathbf{x}$$

Now we come back to functional (1) and suppose that *there is a one-parameter group of transformations of the form*

$$\bar{x} = \varphi(x, \mathbf{y}, t), \quad \bar{\mathbf{y}} = \mathbf{\Phi}(x, \mathbf{y}, t) \tag{26}$$

defined in the $(n+1)$*-dimensional space* $x, \mathbf{y}$ (where $\mathbf{y} = (y_1, y_2, \ldots, y_n)$) *under which this functional remains invariant.* (For the $(n+1)$-dimensional space with points $(x, y_1, \ldots, y_n)$ formula (26) is equivalent to (23) if we put $x = x_1$ and $y_i = x_{i+1}$, $i = 1, 2, \ldots, n$, but it appears more convenient to use (26) because the role of the coordinate x in functional (1) differs from that of the variables y.) The invariance condition means that if we interpret (1) as a functional dependent on the variable curve $\mathbf{y} = \mathbf{y}(x)$ $(a \leqslant x \leqslant b)$,

integral (1) must assume equal values on all the curves going into one another under mapping (26).

In accordance with (24), let us put

$$\varphi_0(x, y) = \varphi_t'(x, y, 0), \quad \Phi_0(x, y) = \Phi_t'(x, y, 0)$$

and compare the value of the functional assumed on an arc of an extremal with its value attained on the curve into which this arc goes under mapping (26) corresponding to an infinitesimal value of t. By the invariance of the functional, the left member of formula (8) is then equal to zero, and the integral on the right-hand side vanishes, and therefore formula (25) yields

$$(\mathbf{p} \cdot \Phi_0 t - H\varphi_0 t) \Big|_{x=a}^{b} = 0$$

that is

$$(\mathbf{p} \cdot \Phi_0 - H\varphi_0) \Big|_{\substack{x=b \\ y=y\,(b)}} = (\mathbf{p} \cdot \Phi_0 - H\varphi_0) \Big|_{\substack{x=a \\ y=y\,(a)}}$$

the latter equality being fulfilled for any two points of the extremal in question. Hence, the relation

$$\mathbf{p} \cdot \Phi_0(x, y) - H(x, y, \mathbf{p})\, \varphi_0(x, y) = \text{const} \tag{27}$$

holds along every extremal, and thus we have obtained a first integral of the system of Euler's equations corresponding to functional (1). Formula (27) expresses the **Noether theorem** for functionals of form (1) (Amalie Emmy Noether (1882-1935), a German mathematician).

For instance, let the function F in functional (1) be independent of x. This means that the functional is invariant with respect to the group of transformations $\bar{x} = x + \alpha$, $\bar{y} = y$. Then we have $\Phi_0 = 0$, $\varphi_0 = 1$, and therefore (27) yields the first integral $H = \text{const}$ which was mentioned in Sec. 2.

A group of transformations under which a functional remains invariant may involve more than one parameter. For example, in the case $n = 3$ we can consider the two-parameter group of motions along the coaxial helices or the three-parameter group of all rotations about a point, etc. (let the reader explain why the last example involves three parameters). Note that the latter group is *noncommutative*: the result of two consecutively performed mappings belonging to the group depends, in the general case, on the order in which the mappings are applied. If there are k parameters we write the formula expressing the mappings of the family in the form

$$\bar{\mathbf{x}} = \Phi(\mathbf{x}, t_1, t_2, \ldots, t_k) \tag{28}$$

instead of (23). Family (28) must contain the identity mapping, the inverse mapping of every mapping belonging to it and the

result of consecutive application of any two mappings. If a functional is invariant relative to this group we can obtain, by means of the differentiation with respect to each parameter, k independent integrals of type (27) for the corresponding system of Euler's equations, which makes it possible to reduce by k the number of the equations in the system.

The theory of continuous groups of transformations which are also called **Lie groups**, after M.S. Lie (1842-1899), a Norwegian mathematician, is an extensively developed division of modern mathematics with many useful applications.

A functional can also be invariant with respect to a group of transformations involving a certain number k of arbitrary functions. (For example, a functional in parametric form (1.54) turns out to be invariant under any change $\bar{t}=\bar{t}(t)$ of the variable of integration because this change simply means that we pass to a new parametrization of the line (L) while the value of the functional is independent of the choice of the parametrization. Consequently, in this case we have $k=1$.) A.E. Noether proved another theorem, applicable to functionals possessing an invariance property of that kind, which states that in the case of k arbitrary functions there are k independent functional relationships in the system of Euler's equations, and therefore k equations can be discarded.

6. The Case of Functions of Several Variables. Here we shall generalize the Noether theorem to functionals dependent on functions of several variables. We shall limit ourselves to functionals of form (1.55) in which we shall write, respectively, x_1 and x_2 instead of x and y, which is more convenient because the resultant formulas hold for any number m of independent variables $x_1, x_2, \ldots, x_m$. To begin with, we write the general formula

$$\delta \int_{(G)} F\left(\mathbf{x}, z \frac{\partial z}{\partial \mathbf{x}}\right) dG =$$

$$= \int_{(G)} (F'_z - \operatorname{div} F'_{\mathbf{p}})\, \delta z\, dG + \int_{(\Gamma)} (F'_{\mathbf{p}})_n\, \delta z\, d\Gamma + \int_{(\Gamma)} F\,(\delta \mathbf{r})_n\, d\Gamma \quad (29)$$

for the variation of functional (1.55) arising when both the function $z(x, y)$ and the boundary (Γ) of the domain of integration (G) are varied. Here, as usual, the symbol $F'_{\mathbf{p}}$ designates the vector with projections F'_{p_i} $(i=1, 2)$, and $\mathbf{r}=\sum x_i \mathbf{e}_i$ is the radius vector. Formula (19) generalizes formula (1.91) and is derived by analogy with the latter. Besides, by analogy with (1.88), we obtain the relation

$$\delta z|_{(\Gamma)} = \delta[z(\mathbf{x})] - \operatorname{grad} z \cdot \delta \mathbf{r}$$

By the way, if we regard the variation of the boundary (Γ) as the result of the motion of its points along the normals to (Γ), formula (29) implies that the expressions $(F'_{\mathbf{p}})_n$ and $F-(F'_{\mathbf{p}})_n z'_n$ must vary continuously as the variable point crosses a line of discontinuity of the function $z = z(x_1, x_2)$. Let the reader prove this assertion which generalizes the Weierstrass-Erdmann conditions (see (1.107)) to piecewise smooth surfaces satisfying the Euler-Ostrogradsky equation and giving stationary values to functional (1.55).

If the variation of the boundary of the domain is regarded as resulting from a variation of all the points of the domain, the quantity $\delta \mathbf{r}$ is defined throughout the domain (G), and therefore we can apply Ostrogradsky's formula $\int\limits_{(\Gamma)} A_n\, d\Gamma = \int\limits_{(G)} \operatorname{div} \mathbf{A}\, dG$ to the line integrals in formula (29), which, together with the above formula for $\delta z|_{(\Gamma)}$, yields

$$\delta \int\limits_{(G)} F\left(\mathbf{x}, z, \frac{\partial z}{\partial \mathbf{x}}\right) dG = \int\limits_{(G)} [(F'_z - \operatorname{div} F'_{\mathbf{p}})\, \delta z + \\ + \operatorname{div} (F'_{\mathbf{p}} (\delta [z(\mathbf{x})] - \operatorname{grad} z \cdot \delta \mathbf{r}) + F \delta \mathbf{r})]\, dG \tag{30}$$

Now suppose that we are given a one-parameter group of transformations of the form

$$\bar{\mathbf{x}} = \varphi(\mathbf{x}, z, t), \qquad \bar{z} = \Phi(\mathbf{x}, z, t) \tag{31}$$

under which the functional under consideration remains invariant in the same sense as before. Then, proceeding from formula (30) and arguing as in deriving equalities (27), we conclude that for any extremal $z = z(\mathbf{x})$ and any domain (G) the relation

$$\int\limits_{(G)} \operatorname{div} [F'_{\mathbf{p}} (\Phi_0 - \operatorname{grad} z \cdot \varphi_0) + F \varphi_0]\, dG = 0$$

must hold. It follows, by the arbitrariness of the domain (G), that

$$\operatorname{div} [F'_{\mathbf{p}} (\Phi_0 - \operatorname{grad} z \cdot \varphi_0) + F \varphi_0] = 0 \tag{32}$$

Relation (32) which holds for every extremal $z(\mathbf{x})$ satisfying the Euler-Ostrogradsky equation is an analogue of first integral (27) for functional (1.55). Note that the symbol $\mathbf{p}$ in relation (32) designates the vector with projections z'_{x_i}, $i = 1, 2$ (while in (27) it has another meaning).

The same formula (32) is obtained for a more general (than (31)) transformation expressed by the formulas

$$\bar{\mathbf{x}} = \varphi(\mathbf{x}, z, z'_{\mathbf{x}}, t), \quad \bar{z} = \Phi(\mathbf{x}, z, z'_{\mathbf{x}}, t)$$

if a functional of type (1.55) is invariant with respect to it. For any fixed t and any function $z(\mathbf{x})$ these formulas express a parametric representation of the surface $\bar{z}(\bar{\mathbf{x}})$ (why?), and thus, like formulas (31), they determine a transformation of surfaces into surfaces (of course, it is required that the conditions

$$\varphi(\mathbf{x}, z, z'_{\mathbf{x}}, 0) \equiv \mathbf{x}, \quad \Phi(\mathbf{x}, z, z'_{\mathbf{x}}, 0) \equiv z$$

be fulfilled).

Formula (32) means that the vector field

$$F'_{\mathbf{p}}(\Phi_0 - \operatorname{grad} z \cdot \varphi_0) + F\varphi_0$$

has no sources of vector lines in the space of vectors $\mathbf{x}$ for every solution $z(\mathbf{x})$ of the Euler-Ostrogradsky equation (the relationship between the divergence of a vector field and the notion of a source of vector lines is the same for any dimension; e. g. for the case $n=3$, see [107], Secs. XVI.23, 24). It thus expresses a law of conservation whose physical significance depends on the concrete type of the functional.

It is also possible to derive a formula similar to (32) which applies to functionals of the form $\int F\left(\mathbf{x}, z, \frac{\partial z}{\partial \mathbf{x}}\right) dG$ dependent on n functions of m variables. In this case the expression $F'_{\mathbf{p}}$ is understood as the $(m \times n)$-matrix with elements $a_{\alpha\beta} = F'_{\partial z_\beta / \partial x_\alpha}$ $(\alpha = 1, \ldots, m;\ \beta = 1, \ldots, n)$ and grad $\mathbf{z}$ as the $(n \times m)$-matrix with elements $b_{ij} = \frac{\partial z_i}{\partial x_j}$ $(i = 1, \ldots, n;\ j = 1, \ldots, m)$.

7. Hamilton-Jacobi Equation. Let us take a functional of form (1). The general solution of the corresponding system of Euler's equations involves $2n$ arbitrary constants, and therefore the specification of two points A and B in the space of variables $x, y_1, y_2, \ldots, y_n$ through which the extremal must pass yields exactly $2n$ finite equations for determining these constants. Therefore, in the general case, there appears a discrete collection of extremals (which may contain one or many extremals or none) joining these points. Let I_{AB} be the value of the functional on each of these extremals, the point A being regarded as its initial point and B as the terminal point.

Let us consider the point A to be fixed while $B(x, y_1, y_2, \ldots, y_n)$ is regarded as the moving point. Then I_{AB} becomes a function of $x, y_1, \ldots, y_n$, and we can write

$$I_{AB} = S(x, y_1, y_2, \ldots, y_n) \tag{33}$$

where S is a function of $n+1$ variables $x, y_1, \ldots, y_n$. Generally speaking, the function S, defined for all the points $B(x, y_1, y_2, \ldots, y_n)$

which can be connected with the point A by an extremal, is multiple-valued. Let us consider a one-valued branch of this function (which we again denote by S). If the point B changes its position, formula (8) implies the relation

$$dS = -H\,dx + \sum_{i=1}^{n} p_i dy_i$$

In other words, we have

$$\frac{\partial S}{\partial x} = -H, \quad \frac{\partial S}{\partial y_i} = p_i \qquad (i = 1, 2, \ldots, n)$$

It follows that function (33) satisfies the partial differential equation of the first order

$$\frac{\partial S}{\partial x} + H\left(x, y_1, y_2, \ldots, y_n, \frac{\partial S}{\partial y_1}, \frac{\partial S}{\partial y_2}, \ldots, \frac{\partial S}{\partial y_n}\right) = 0 \qquad (34)$$

known as the **Hamilton-Jacobi equation.**

It can easily be proved that *if* $S = \tilde{S}(x, y_1, y_2, \ldots, y_n, \alpha)$ *is an arbitrary one-parameter family of solutions of equation* (34), *every extremal must satisfy the relation*

$$\tilde{S}'_\alpha(x, y_1, y_2, \ldots, y_n, \alpha) = \beta \qquad (35)$$

where $\beta = const$. Indeed, using equations (7) we obtain the equality

$$\frac{d\tilde{S}'_\alpha}{dx} = \frac{\partial \tilde{S}'_\alpha}{\partial x} + \sum_i \frac{\partial \tilde{S}'_\alpha}{\partial y_i}\frac{dy_i}{dx} = \frac{\partial^2 \tilde{S}}{\partial x \partial \alpha} + \sum_i \frac{\partial H}{\partial p_i}\frac{\partial^2 \tilde{S}}{\partial y_i \partial \alpha}$$

where the rightmost expression is the result of the differentiation of the left-hand side of (34) (into which $\tilde{S}$ is substituted for S) with respect to α, and therefore $\frac{d\tilde{S}'_\alpha}{dx}$ is equal to zero whence follows (35). Relation (35) makes it possible to eliminate one of the unknown functions in the system of Euler's equations (2) and thus to reduce by 2 the order of the system in the sense that this system is reducible to a system of $2n-2$ first-order equations.

Similarly, knowing a k-parameter family of solutions of equation (34) we can reduce the order of the system of Euler's equations (2) by $2k$ with the aid of the relations

$$\tilde{S}'_{\alpha_i}(x, y_1, \ldots, y_n, \alpha_1, \ldots, \alpha_k) = \beta_i \quad (i = 1, 2, \ldots, k) \qquad (36)$$

where β_i are constants. If $k = n$, relations (36) represent the general solution of the system of Euler's equations (2).

Now, let the Hamiltonian function H be independent of x; then, when solving equation (34), we can confine ourselves to functions of the form

$$S = -hx + W(y_1, y_2, \ldots, y_n), \quad h = \text{const}$$

and rewrite equation (34) in the form

$$H\left(y_1, y_2, \ldots, y_n, \frac{\partial W}{\partial y_1}, \frac{\partial W}{\partial y_2}, \ldots, \frac{\partial W}{\partial y_n}\right) = h$$

Suppose that we have managed to construct a function $W(y, P)$ $(y=(y_1, \ldots, y_n), P=(P_1, \ldots, P_n))$ involving n parameters P such that if it is substituted into the left-hand side (of the last equation) the latter becomes equal to a constant $h=h(P)$. We can choose W as the generating function of a canonical transformation (see Sec. 3), i.e. put

$$p_i = \frac{\partial W}{\partial y_i}, \quad Y_i = \frac{\partial W}{\partial P_i}, \quad \tilde{H} = H(y, p) = h(P)$$

Then, in the new variables, the canonical system of equations turns into

$$\frac{dP_i}{dx} = 0, \quad \frac{dY_i}{dx} = \frac{\partial h}{\partial P_i}$$

and therefore the general solution of the system has the form

$$P_i = C_i, \quad Y_i = h'_{P_i}(C_1, C_2, \ldots, C_n)\, x + D_i \quad (i = 1, 2, \ldots, n)$$

where C_i and D_i $(i=1, 2, \ldots, n)$ are arbitrary constants.

The quantity I_{AB} introduced in the first paragraph of this section admits of an interesting geometric interpretation. For the sake of convenience, let us regard the coordinates as being equivalent and pass to parametric form (1.54) of the functional. Then, in the case of n variables $x_1, x_2, \ldots, x_n$, the functional takes the form

$$I\{(L)\} = \int_{t_{\text{initial}}}^{t_{\text{terminal}}} F_1(x_1, x_2, \ldots, x_n, \dot{x}_1, \dot{x}_2, \ldots, \dot{x}_n)\, dt \qquad (37)$$

where F_1 is a homogeneous function of degree 1 with respect to the derivatives $\dot{x}_1, \dot{x}_2, \ldots, \dot{x}_n$. We shall also assume that $t_{\text{initial}} < t_{\text{terminal}}$.

Suppose that $F_1 > 0$ (of course, except the values $\dot{x}_1 = \dot{x}_2 = \ldots = \dot{x}_n = 0$ for which, by the homogeneity property, we have $F_1 = 0$) and that for any two points A and B functional (37) attains a minimum value on the family of curves joining these points. Denoting this minimum value by $\rho(A, B)$ we can assert (let the reader prove it!) that the quantity ρ possesses the following properties:

1. $\rho(A, B) > 0$ for $A \neq B$, and $\rho(A, A) = 0$;
2. $\rho(A, B) = \rho(B, A)$;
3. $\rho(A, C) \leqslant \rho(A, B) + \rho(B, C)$ (this is the *triangle inequality*).

These three properties are known as the *axioms of metric space.* If they hold, the quantity $\rho(A, B)$ can be taken as the *generalized distance* between the points A and B. A space for which the notion of a generalized distance is introduced is referred to as a **metric space,** and $\rho(A, B)$ is called a **distance function** (or **metric**). The value of functional (37) is then naturally interpreted as the (generalized) length of the curve (L) relative to the chosen metric, extremals are understood as geodesic lines of the metric (see Sec. 2.6) and the values of I_{AB} as the lengths of the geodesics connecting the points A and B. As is seen from (37), the linear element (the element of arc length) is expressed by the formula

$$dI = F_1(x_1, x_2, \ldots, x_n, dx_1, dx_2, \ldots, dx_n) \qquad (38)$$

which in many cases makes it possible to pass to the ordinary form of a metric tensor in a Riemannian space (see Sec. V.2.5).

The space in which the above construction has been realized must not necessarily be Euclidean; it can be an arbitrary manifold with generalized coordinates $x_1, x_2, \ldots, x_n$. A functional of form (37) defined in this manifold and satisfying the above axioms turns it into a metric space. If the square of the right member of formula (38) is a quadratic form in the differentials $dx_1, dx_2, \ldots, dx_n$ this metric space is a Riemannian space.

The above requirements by far not always hold. In particular, there may exist pairs of points that cannot be joined by a line on which the functional assumes a real value, and then the notion of a generalized distance between the points cannot be introduced in a meaningful fashion. For instance, this is the case for the functional

$$\int \sqrt{dx_1^2 - dx_2^2} \text{ if } |x_{1B} - x_{1A}| < |x_{2B} - x_{2A}| \text{ (why?)}$$

8. Lobachevskian Plane. Here we shall discuss a remarkable example. Consider the functional

$$I\{(L)\} = \int_{(L)} = \frac{1}{y}\sqrt{\dot{x}^2 + \dot{y}^2}\, dt$$

as being defined on the curves $x = x(t)$, $y = y(t)$ lying in the upper half-plane $-\infty < x < \infty$, $0 < y < \infty$. By virtue of Sec. 1.17, the extremals of this functional can be thought of as rays of light propagating in the half-plane, the velocity of light at each point (x, y) being proportional to y. In the metric corresponding to this functional the square of the element of arc length is written in the form

$$dI^2 = \frac{1}{y^2}(dx^2 + dy^2) \qquad (39)$$

Integrating Euler's equations corresponding to the functional we find (check it up!) that the extremals, that is the geodesic lines (relative to the new metric), are the semicircular arcs with centres on the x-axis lying in the upper half-plane and also their limiting forms which are the half-lines $x = \text{const}$, $y > 0$. If we conditionally call these extremals "straight lines" and the upper half-plane the "plane" while the term "point" is used in the ordinary sense, it can readily be shown that all the axioms (postulates) of plane Euclidean geometry, except *Euclid's parallel postulate*, hold. (As is known, Euclid's parallel postulate reads as follows: through a point not on a straight line there is no more than one line parallel to that straight line.) For instance, let the

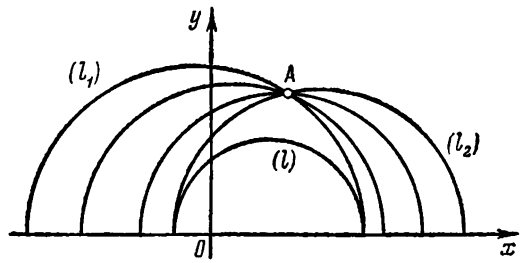


Fig. 104

reader verify that there is exactly one "straight line" passing through any two points. At the same time, through the point A (shown in Fig. 104) not lying on the "straight line" l we can draw infinitely many "straight lines" which do not meet (l) when "infinitely extended". (It should be noted that, according to formula (39), a point moving in a "straight line" and approaching the x-axis "travels to infinity" relative to the new metric.) These "straight lines" through the point A "parallel to (l)" form a "pencil of straight lines" bounded by the two "straight lines" (l_1) and (l_2) (Fig. 104) one of which asymptotically approaches (in the new metric, of course) the "straight line" (l) from the right while the other from the left.

A geometry (understood as a collection of "points", "lines", etc. with certain relationships between them specified by a system of axioms) in which all the postulates of Euclidean geometry hold except Euclid's parallel postulate is called the **Lobachevskian geometry**. It was independently discovered in the 19th century by J. Bolyai (1802-1860), a prominent Hungarian mathematician, the great German mathematician K.F. Gauss and the great Russian mathematician N.I. Lobachevsky (1792-1856). Lobachevsky developed this new geometry most extensively and was the first to publish the report announcing this discovery.

Many interesting results of the Lobachevskian planimetry are easily obtained on the basis of the above realization of the **Lobachevskian plane**. The right-hand side of formula (39) being *isotropic* (in the sense that it is independent of the direction determined by the ratio $dx:dy$), it follows that in the Lobachevskian geometry the angle between two lines at the point of their intersection coincides with the ordinary angle between these lines in the xy-plane. This implies that a "motion" of the Lobachevskian plane should be understood as a conformal mapping of the upper half-plane into itself (see Sec. II.1.8), the latter being a fractional linear mapping which not only preserves the angles but also transforms "straight lines" into "straight lines" (according to the circular property of fractional linear mappings). Formula (39) yields the following formula for the "area" of an arbitrary figure (S):

$$S=\iint\limits_{(S)}\frac{1}{y^2}\,dx\,dy$$

Integrating with respect to y we obtain another formula of "area":

$$S=\oint\limits_{(L)}\frac{1}{y}\,dx \tag{40}$$

Here (L) is the contour bounding the figure (S), and it is described in the positive direction.

Now consider the "triangle" ABC with angles α, β and γ shown in Fig. 105. The arc AB can be represented parametrically in the form $x=a+R\cos\varphi$, $y=R\sin\varphi$, and, consequently, the contribution made by this arc to integral (40) is equal to

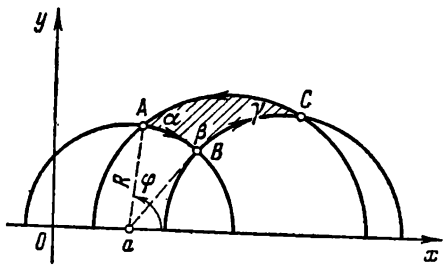


Fig. 105

$$\int\limits_{AB}\frac{1}{R\sin\varphi}\,d\,(a+R\cos\varphi)=\varphi_{AB}$$

that is to the angle at which the arc AB is seen from its centre. We similarly find the contributions due to the other two arcs, and thus obtain the formula

$$S=\varphi_{AB}+\varphi_{BC}-\varphi_{AC} \tag{41}$$

But if the closed contour ABC is described in the positive direction its tangent line turns through the angle $-\varphi_{AB}+(\pi-\beta)-\varphi_{BC}+(\pi-\gamma)+\varphi_{AC}+(\pi-\alpha)$ which must be equal to 2π. Hence,

we finally derive from formula (41) the relation

$$S=\pi-(\alpha+\beta+\gamma)$$

We see that in the Lobachevskian geometry the sum of the interior angles of a triangle is always less than π. We also see that the area of a triangle depends solely on the sum of its angles; if the area decreases the sum of the angles becomes closer to π. Besides, it is obvious that the notion of similarity of geometric objects is inapplicable to the Lobachevskian geometry because if the corresponding angles of two "triangles" are equal the "triangles" themselves are equal (prove this property!).

To interpret the Lobachevskian stereometry we can make a similar construction in the half-space $z>0$ of the three-dimensional space with coordinates x, y, z (the realization of this construction is left to the reader).

9. Variational Principles. Every *variational principle* states that, *given the conditions of a problem belonging to a certain class of problems, the actual state, process, etc., compatible with the constraints imposed by the conditions of the problem, differs from all the other states, processes, etc., permitted by the constraints, in that it gives to a functional, characteristic of this class of problems, a stationary value.* For instance, a variational principle of mechanics indicates, in the above sense, in what way the actual trajectory of motion (the **natural trajectory**) differs from all the other kinematically possible trajectories and the like.

Thus, a variational principle is characterized by a class of problems, compatibility conditions for the set of admissible processes (for definiteness, we speak of processes) with constraints included into the formulation of the problem and by a functional defined for these processes whose values are compared and which must assume a stationary value. We sometimes also speak of extremal values of the functional, which is close to the above definition but not completely equivalent to it.

In the introduction to the present chapter we mentioned the importance of variational principles and problem solving methods which they provide. Every such principle belongs to the division of science which treats of the class of problems the principle applies to and is not therefore studied in general mathematical courses. But here we shall discuss some of the principles which are of general character or are closely related to mathematics or provide useful examples of application of mathematical techniques.

We shall roughly classify variational principles into two groups. The first group deals with problems of "pure" and applied physics and other disciplines. It is this group of principles to which the definition given at the beginning of this section literally applies. These principles (for instance, *Fermat's principle* in optics) are

used for studying real physical processes. A principle of this kind can be derived (irrespective of its real historical origin) from the theory of the process (for example, from the differential equation of the process which turns out to be Euler's equation for a functional that can be chosen as a characteristic of the process). In the problems of this type we usually speak of a stationary value of the functional because extremality is inessential for the local properties of the process described by the differential equation, although there are some classes of problems in which extremality provides some additional information (for instance, it can be connected with the stability of the process and the like). It should be noted that even in the case when a variational principle is completely equivalent to the corresponding differential equation it nevertheless may be extremely useful because of its generality (compared with the differential equation) which makes it possible to establish some important analogies and generalizations and also provides for the application of numerical methods.

The other group of variational principles has been developed in recent decades and deals with problems belonging to regulation theory, mathematical economics and some other related sciences. In these problems we speak of working out the best strategy, behaviour, which guarantees the maximum "profit". For this purpose we introduce a functional, called a cost function (functional) or objective function (see Sec. IV.5.1), defined for an admissible class of strategies whose value must be maximized. When referring to problems of that group we usually speak of **optimal control problems**. Here we shall not deal with these problems and only mention that a feature of this class of problems is that they often include unilateral constraints (Sec. 1.19) and therefore their solutions do not necessarily give stationary values to the corresponding functionals. This makes it necessary to elaborate some special methods differing from classical methods of the calculus of variations which most often use stationary values.

One of the simplest and most visual principles belonging to the first group is **Fermat's principle** of geometrical optics (e.g. see [107], Sec. IV.19). The elementary formulation of this principle states that from all the paths connecting one point with another the ray of light chooses the natural trajectory which it traverses in the minimum time. Here we compare all the possible paths compatible with the construction of the optical device under consideration, and the functional to be minimized is the time in which the ray passes from one point to another. Fermat's principle implies the laws of light propagation in homogeneous and nonhomogeneous media and the laws of reflection and refraction. A more thorough analysis shows that there are cases when the

actual time has a minimax but not a minimum value (e.g. see [140], § XII.10). Thus, a more precise formulation of Fermat's principle involves a stationary value of time. The local form of Fermat's principle can be derived from the wave theory of light (there are also cases when an actual trajectory of light does not correspond to a stationary value of time; these cases are connected with the phenomena of interference of light). In some simpler situations the minimality of time of light propagation is implied by the particular structure of the functional which must attain its stationary values. (For example, let the reader prove, by passing to the integration with respect to x, that every sufficiently small arc of an extremal gives functional (1.95) a strong minimum, and even the absolute minimum, among the class of all curves connecting the end points of that small arc.)

10. Hamilton's Principle. The Simplest Case. We shall begin with a simple example demonstrating *Hamilton's variational principle* which is one of the most general principles of modern natural science. Suppose that a material point of mass m can be in motion along the x-axis, the external force acting upon the point being also in the direction of the x-axis. Let this force depend on the coordinate x and on the time of motion t and be expressed by a given function $f(x, t)$. As is known, in this case the law of motion $x = x(t)$ must satisfy the differential equation

$$m\ddot{x} = f(x, t) \tag{42}$$

implied by Newton's second law of dynamics. Let us denote by $U(x, t)$ an antiderivative (with respect to the variable x) of the function $-f(x, t)$. For instance, we can put

$$U(x, t) = -\int_{x_0}^{x} f(s, t)\, ds$$

The function $U(x, t)$ is the **potential** (see Sec. I.2.1) of the given field of force. We also need the expression

$$T = \frac{m\dot{x}^2}{2}$$

which is the **kinetic energy** of the mass point. It is obvious that equation (42) can be rewritten in the form

$$-U'_x - \frac{d}{dt} T'_{\dot{x}} = 0 \tag{43}$$

Furthermore, putting

$$L(t, x, \dot{x}) = T - U$$

(the function L is the **Lagrangian function** or the **kinetic poten-**

tial) and taking into account that U is independent of $\dot{x}$ and T of x we can write equation (43) as

$$L'_x - \frac{d}{dt} L'_{\dot{x}} = 0 \tag{44}$$

We see that (44) is Euler's equation corresponding to the functional

$$\int_{t_0}^{t_1} L(t, x, \dot{x})\, dt \tag{45}$$

Functional (45) is known as the **(Hamiltonian) action**. Condition (44) means that the actual law of motion $x(t)$ gives a stationary value to functional (45) for fixed end point values

$$x(t_0) = x_0, \quad x(t_1) = x_1$$

Hence, the stationarity of the value of the functional is completely equivalent to equation (42), and consequently, Newton's differential law of motion can be formulated as follows: *given initial and terminal states of the system* (that is the initial and terminal time moments and the corresponding values of the coordinates of the material point), *the actual law of motion is the one for which the action takes on a stationary value.* This is **Hamilton's principle** for the simplest situation we have considered here.

The structure of the integrand in formula (45) makes it possible to prove (see the beginning of Sec. 2.6) that for every sufficiently small interval of time the actual law of motion not only gives the action a stationary value but also a minimum. On the other hand, there are simple examples indicating that, generally speaking, for long time intervals the action may assume a minimax value. Therefore, in the general formulation of Hamilton's principle we speak of stationary values.

Let us consider an example of a linear oscillator. Here we have $f = -kx$ where k is the constant stiffness factor, and, consequently, $U = \frac{kx^2}{2}$ and

$$L = \frac{1}{2}(m\dot{x}^2 - kx^2)$$

Thus, integral (45) turns into a quadratic functional with constant coefficients. In Sec. 2.4 we investigated functionals of that type. In terms of the notation of the present example the results of Sec. 2.4 state that for $t_1 - t_0 < \pi\sqrt{\frac{m}{k}}$ the functional has a minimum and for $t_1 - t_0 > \pi\sqrt{\frac{m}{k}}$ a minimax. Note that

$\pi\sqrt{\frac{m}{k}}=\frac{\pi}{\omega_0}$ is the half-period of free vibrations of the oscillator, that is the time interval between two successive moments at which the equilibrium state $x=0$ is passed.

We see that here the deduction of the variational principle reduces to representing equation of motion (42) in the form of Euler's equation (44). But nevertheless this simple transformation proves useful for further generalizations; besides, as will be seen in § 4, there are a number of other one-dimensional problems which are more conveniently solved after a transformation of that kind.

In this connection we can pose the following general question: what are the conditions under which a given differential equation is representable as Euler's equation for a functional? Let us limit ourselves to functional (1.18). Then, according to (1.37), we must consider a differential equation of the form

$$U(x, y, y')y''+V(x, y, y')=0 \tag{46}$$

Putting $y'=z$ we see that from the equalities

$$\left.\begin{aligned} &U(x, y, z)\equiv -F''_{zz}(x, y, z)\\ &V(x, y, z)\equiv F'_y(x, y, z)-F''_{xz}(x, y, z)-F''_{yz}(x, y, z)\,z \end{aligned}\right\} \tag{47}$$

it follows that for the choice of the function F to be possible it is necessary that

$$V'_z\equiv U'_x+U'_y z \tag{48}$$

It can also be shown that this condition is sufficient. To construct this function we find F, to within a linear term with respect to z, from the first equality (47) and then determine this term by substituting the expression thus obtained into the second equality (47). Theoretically, we can always multiply both sides of original equation (46) by an appropriately chosen factor $R(x, y, y')$ such that the resultant equation satisfies condition (48), but this can prove impractical. If condition (48) is fulfilled the function F is determined to within a summand which is obtained if we put $U\equiv V\equiv 0$ in equalities (47) (why?). The first relation (47) then yields

$$F=a(x, y)z+b(x, y)$$

Substituting this expression into the second relation (47) we see that

$$a'_x\equiv b'_y, \text{ i.e. } a=\varphi'(x, y) \text{ and } b=\varphi'_x(x, y)$$

Thus, the above-mentioned linear (with respect to z) summand in F is of the form

$$\varphi'_y(x, y)y'+\varphi'_z(x, y)=\frac{d}{dx}\varphi(x, y)$$

This result is quite obvious: if, for instance, the functional $I\{y\}$ is considered under boundary conditions (1.26), the addition of a term of that type to the function F simply leads to the appearance of an additional constant term in $I\{y\}$.

The results established in the last paragraph are immediately extended to functionals of form (45) dependent on vector functions $y(x)$ and also to functionals of functions of several variables. For example, the Euler-Ostrogradsky equation (1.59) remains invariant if we add to the function F an arbitrary term of the form

$$\frac{\partial}{\partial x}A(x, y, z)+\frac{\partial}{\partial y}B(x, y, z)$$

(referred to as a **divergence expression**); let the reader check up this assertion.

11. Hamilton's Principle for Systems with Finite Number of Degrees of Freedom. Now let us take a system of differential equations of the form

$$m_i\ddot{x}_i=f_i(x_1, x_2, \ldots, x_N, t) \qquad (i=1, 2, \ldots, N)$$

which generalizes equation (42) to the case of an arbitrary number N ($N\geqslant 1$) of degrees of freedom. For instance, equations of this type may describe the motion of a system of r material points in space which are not subject to any constraints and are acted upon by forces dependent on the coordinates of the points and on time. (In this interpretation x_1, x_2, x_3 are the Cartesian coordinates of the first point whose mass is $m_1=m_2=m_3$; x_4, x_5, x_6 are the coordinates of the second point with mass $m_4=m_5=m_6$, etc. The number N of degrees of freedom of this system is equal to $3r$.) Suppose that the field of force acting upon the system has a potential $U(x_1, x_2, \ldots, x_N, t)$, that is we have

$$f_i=-\frac{\partial U}{\partial x_i} \qquad (i=1, 2, \ldots, N) \tag{49}$$

In contrast to Sec. 10, this potential (**potential energy** of the system) by far not always exists if $N\geqslant 2$. For the potential to exist it is necessary and sufficient that the equalities

$$\frac{\partial f_i}{\partial x_j}\equiv\frac{\partial f_j}{\partial x_i} \qquad (i, j=1, 2, \ldots, N)$$

hold (cf. Sec. I.2.1). Reasoning like in Sec. I.2.1, we can easily show that U is equal to the work performed by the field of force at a fixed time moment t when the system is made to pass from any current state $(x_1, x_2, \ldots, x_N)$ to a given state $(x_{10}, x_{20}, \ldots, x_{N0})$; the work must not depend on the way in which the system moves (this is the criterion for the system to be potential). The arbitrariness in the choice of the terminal state $(x_{10}, x_{20}, \ldots, x_{N0})$

is equivalent to the possibility of adding an arbitrary function of time to U, which is inessential for our further considerations.

Now, putting

$$T=\sum_i \frac{1}{2} m_i \dot{x}_i^2,\ L(t, x_1, x_2, \ldots, x_N, \dot{x}_1, \dot{x}_2, \ldots, \dot{x}_N)=T-U$$

and reasoning by analogy with Sec. 10 we arrive at the same conclusion corcerning the stationarity of the action, that is of the integral

$$\int_{t_0}^{t_1} L(t, x_1, x_2, \ldots, x_N, \dot{x}_1, \dot{x}_2, \ldots, \dot{x}_N)\, dt$$

for the actual law of motion of the system.

It is proved in mechanics that Hamilton's principle also remains valid for the case when the system of material points is subject to holonomic constraints of the form

$$g_i(x_1, x_2, \ldots, x_N, t)=0 \quad (i=1, 2, \ldots, \nu, \quad \nu<N) \tag{50}$$

In the latter case, when applying Hamilton's principle, we must only compare the laws of motion which are compatible with the constraints, that is satisfy equations (50). In other words, in this case we deal with a *conditional stationary value* of the action (see Sec. 1.13). To pass to unconditional stationary values (which sometimes is more convenient) we introduce **generalized coordinates** $q_1, q_2, \ldots, q_n$ (where $n=N-\nu$ is the number of degrees of freedom of the system) specifying the state of the system. The coordinates of the points of the system are expressed in terms of the generalized coordinates by means of relations

$$x_i=x_i(q_1, q_2, \ldots, q_n, t) \quad (i=1, 2, \ldots, N) \tag{51}$$

(which are nothing but a parametric representation of the general solution of system of equations (50)), and therefore the potential U, the kinetic energy T and the Lagrangian function L are expressed as functions of $t, q_1, \ldots, q_n, \dot{q}_1, \ldots, \dot{q}_n$: $U=U(t, q_1, q_2, \ldots, q_n)$, $T=T(t, q_1, q_2, \ldots, q_n, \dot{q}_1, \dot{q}_2, \ldots, \dot{q}_n)$, $L=T-U=L(t, q_1, q_2, \ldots, q_n, \dot{q}_1, \dot{q}_2, \ldots, \dot{q}_n)$. The way in which the generalized coordinates have been introduced automatically takes into account the constraints imposed on the system, and, consequently, we arrive at the problem of determining an *unconditional stationary value* of the functional

$$\int_{t_0}^{t_1} L(t, q_1, q_2, \ldots, q_n, \dot{q}_1, \dot{q}_2, \ldots, \dot{q}_n)\, dt \tag{52}$$

for given initial and terminal states of the system.

The function T (and therefore L as well) is a polynomial of the second degree with respect to $\dot{q}_1, \dot{q}_2, \ldots, \dot{q}_n$ whose coefficients are dependent on $t, q_1, \ldots, q_n$, and the quadratic form composed of its second-degree terms is positive definite (why?). But in all other respects the function L can be quite arbitrary, and thus, the methods of investigation of mechanical systems with finite number of degrees of freedom considered here and the results obtained are applicable to any functional (1) whose integrand function has the structure indicated above. From the beginning of Sec. 2.6 it follows that the actual law of motion which is an extremal of functional (52) gives a strong minimum to the functional for every sufficiently small time interval.

If we have an *autonomous system* (whose potential U and constraints (50) do not involve t) the right-hand sides of equalities (51) do not depend on t either, and therefore T is a quadratic form in the variables $\dot{q}_1, \dot{q}_2, \ldots, \dot{q}_n$. Then we can apply Euler's theorem on homogeneous functions (e. g. see [107], Sec. IX.12) to the Hamiltonian function (formula (5)), which results in the equality

$$H = -L + \sum \dot{q}_i p_i = -(T-U) + \sum \dot{q}_i T'_{\dot{q}_i} = -(T-U) + 2T = T + U$$

This means that H is the total energy of the system. In the autonomous case the function L under the sign of integration in (52) is independent of t, and hence, by virtue of Sec. 2, we conclude that the function H retains a constant value for any actual law of motion of the system, which provides the first integral $H = \text{const}$. Consequently, the total energy of an autonomous Hamiltonian (canonical) system remains constant in the process of motion. Therefore these systems (for which the **law of conservation of energy** holds) are referred to as **conservative systems.**

The canonical variables $p_i = T'_{\dot{q}_i}$ are called **generalized momenta** (for Cartesian coordinates they coincide with ordinary momenta $m_i \dot{x}_i$).

In some cases first integrals can be found by means of the Noether theorem (see Sec. 5; examples are given in [48]). For instance, let the reader use this theorem for proving that if the function L is independent of a coordinate q_i we have the first integral $p_i = \text{const}$ (which expresses the **law of conservation of momentum**). Also derive this result as a direct consequence of Euler's equations.

In the case of autonomous systems Hamilton's principle has an interesting geometric interpretation. We showed that for an autonomous Hamiltonian system we have the first integral

$$T(q, \dot{q}) + U(q) = h = \text{const} \qquad (53)$$

(for brevity, we denote by q the collection of the coordinates $q_1, q_2, \ldots, q_n$). Therefore, we can limit ourselves to considering the laws of motion satisfying the differential constraint (53) with a given value of h. But we have

$$2T = (T-U)+(T+U) = (T-U)+h$$

and, consequently, Hamilton's principle reduces to the form

$$\delta \int T(q, \dot{q})\, dt = 0 \tag{54}$$

On the other hand, we have

$$T = \sum_{i,j} a_{ij}(q)\, \dot{q}_i \dot{q}_j \text{ whence } dt = \frac{1}{\sqrt{T}} \sqrt{\sum_{i,j} a_{ij}(q)\, dq_i\, dq_j}$$

Substituting this expression into (54) and taking into account that $T = h-U$ we receive

$$\delta \int \sqrt{[h-U(q)] \sum_{i,j} a_{ij}(q)\, dq_i\, dq_j} = 0 \tag{55}$$

Thus, we have completely eliminated time t and reduced Hamilton's principle to form (55) considered without constraint (53); given a trajectory of motion of the system in the space of coordinates q, we can find dt with the aid of the constraint and thus obtain the law of motion of the system over this trajectory.

Now let us introduce a metric in space q (the **state space** of the system) by putting

$$ds^2 = [h-U(q)] \sum_{i,j=1}^{n} a_{ij}(q)\, dq_i\, dq_j \tag{56}$$

Then equation (55) can be rewritten as $\delta \int ds = 0$, and thus we see (Sec. 2.6) that the natural trajectories of motion $q = q(t)$ are the geodesic lines of this metric. To different potential fields and different values of the total energy h there correspond different ways of introducing metric (56) in the state space of the system.

12. Hamilton's Principle for Continuous Media. Vibrations of a String. The formulation of Hamilton's principle is irrelevant of the coordinates of the system and is expressed solely in terms of its potential and kinetic energies. This remarkable feature makes it possible to extend the principle to continuous media by applying an appropriate passage to the limit. Hamilton's principle can be generalized to various physical fields possessing natural analogues of these kinds of energy (for instance, to electromagnetic fields, etc.) and also applies to relativistic mechanics.

Here we shall consider some examples of extension of Hamilton's principle. We begin with deriving the equation of small plane transverse vibrations of a taut string of finite length. Let the ends of the string be at the points $x=0$ and $x=l$ of the x-axis, the tension of the string being P. Besides, we shall suppose that the string is acted upon by a transverse force distributed with density $f(x, t)$ along the x-axis. (In a more general approach a *string* is simply understood as a one-dimensional continuous medium without flexing resistance.)

We shall regard the force of tension P as constant in the process of vibration, which is legitimate when the amplitude of vibration is small, and suppose that the displacement of every point x of the string in the vibration process is perpendicular to the x-axis. Let us denote by $u(x, t)$ the transverse deflection of the point x at moment t from the equilibrium state. The function $u(x, t)$ determines the law of vibrations, and its graph plotted against the axes of x and u represents the shape of the string at every fixed time moment t.

Now imagine that the string passes, at moment t, from the state characterized by the function $u(x, t)$ to the state of equilibrium $u \equiv 0$. Then the external and internal forces perform the work

$$P\left(\int_0^l \sqrt{1+u_x'^2}\,dx - l\right) - \int_0^l f(x, t)\,u\,dx$$

(why?) which should be taken as the potential energy U of the system in question. Expanding the radical by Taylor's formula and discarding the terms involving $u_x'^4$ (on condition that $|u|$ is considered to be small) we obtain the expression

$$U = \int_0^l \left[\frac{P}{2}u_x'^2 - f(x, t)\,u\right] dx$$

where the terms in the square brackets are regarded as being of the same order of smallness. The kinetic energy of the string is equal to

$$T = \int \frac{dm \cdot \dot{u}^2}{2} = \int_0^l \frac{\rho}{2} u_t'^2\,dx$$

where ρ is the linear mass density of the string. Therefore, we have

$$T - U = \int_0^l \left[\frac{\rho}{2}u_t'^2 - \frac{P}{2}u_x'^2 + f(x, t)\,u\right] dx \qquad (57)$$

In this problem, instead of integral (45) expressing the action

in the simplest case, we must take the *Ostrogradsky-Hamilton integral*

$$I\{u\}=\int_{t_0}^{t_1} dt \int_0^l \left[\frac{\rho}{2}u_t'^2-\frac{P}{2}u_x'^2+f(x,\ t)\,u\right] dx \tag{58}$$

Since the action must assume a stationary value the function $u(x,\ t)$ is a solution of the corresponding Euler-Ostrogradsky equation (of form (1.59)):

$$f(x,\ t)-\frac{\partial}{\partial t}(\rho u_t')-\frac{\partial}{\partial x}(-Pu_x')=0 \tag{59}$$

Assuming that ρ is constant (a homogeneous string) and putting

$$a=\sqrt{\frac{P}{\rho}} \quad \text{and} \quad f_1(x,\ t)=\frac{1}{\rho}f(x,\ t)$$

we finally derive from (59) the following *equation of vibrations of the string*:

$$u_{tt}''=a^2u_{xx}''+f_1(x,\ t) \tag{60}$$

Formula (57) indicates that here it is natural to speak of the **density of the Lagrange function**

$$\Lambda=\frac{\rho}{2}u_t'^2-\frac{P}{2}u_x'^2+f(x,\ t)\,u$$

Following the pattern of Sec. 1.12 we can replace the string by a system of mass points with finite number of degrees of freedom, which suggests that the quantity

$$\Pi=\frac{\partial\Lambda}{\partial u_t'}=\rho u_t'$$

should be interpreted as the **momentum density**, while the function

$$\eta=-\Lambda+u_t'\Pi=\frac{\rho}{2}u_t'^2+\frac{P}{2}u_x'^2-f(x,\ t)\,u$$

is the **density of the Hamiltonian function** H; the function H is found from the density by means of integration:

$$H=\int_0^l \left[\frac{1}{2\rho}\Pi^2+\frac{P}{2}u_x'^2-f(x,\ t)\,u\right] dx \tag{61}$$

The results of Sec. 11 imply that in the autonomous case, i.e. when f does not depend on t, the function H is constant for every solution of equation (60) *(the law of conservation of total energy)*. This result can also be directly derived from equation (60). Indeed, differentiating integral (61) with respect to the parame-

ter t and taking advantage of the boundary conditions $u|_{x=0} = u|_{x=l} = 0$ we obtain

$$\frac{dH}{dt} = \int_0^l (\rho u_t' u_{tt}'' + P u_x' u_{xt}'' - f_t' u - f u_t') \, dx =$$

$$= \int_0^l \left[\rho u_t' \left(\frac{P}{\rho} u_{xx}'' + \frac{f}{\rho} \right) + P u_x' u_{xt}'' - f_t' u - f u_t' \right] dx =$$

$$= \int_0^l [P (u_t' u_x')_x' - f_t' u] \, dx = -\int_0^l f_t' u \, dx$$

whence $H = \text{const}$ if $f_t' \equiv 0$, which is what we set out to prove.

It should be noted that, in contrast to Sec. 10, the action expressed by formula (58) assumes only minimax values for any arbitrarily small time interval. To elucidate this property let us put, for simplicity, $f \equiv 0$ and consider the function

$$u_0(x, t) = \begin{cases} \sin \frac{k\pi}{l} x \sin \frac{ak\pi}{l} t & \text{for } 0 \leqslant t \leqslant \frac{l}{ak} \\ 0 & \text{for } t > \frac{l}{ak} \end{cases}$$

where k is a fixed integer ($k = 1, 2, 3, \ldots$). For $t_0 = 0$ and an arbitrarily small and fixed value $t_1 > 0$ we can always choose k such that $\frac{l}{ak} < t_1$. Furthermore, it is readily verified that $I\{u_0\} = 0$ and that the Weierstrass-Erdmann conditions (see Sec. 1.20) do not hold on the line $t = \frac{l}{ak}$ along which the smoothness of the surface $u = u_0(x, t)$ is violated. This confirms the above assertion concerning the minimax stationary values (why?).

We suggest that the reader derive the *equation of transverse vibrations of a taut homogeneous membrane*

$$u_{tt}'' = a^2 (u_{xx}'' + u_{yy}'') + f_1(x, y, t) \tag{62}$$

where $a^2 = \frac{P}{\rho}$, $\rho = \text{const}$ is the areal mass density, P is the tension per unit length, $f_1 = \frac{f}{\rho}$, $f = f(x, y, t)$ is the transverse external force per unit area and $u = u(x, y, t)$ is the (small) transverse deflection of the point (x, y) of the membrane at time moment t from the equilibrium state $u \equiv 0$. (In a more general treatment a *membrane* is simply understood as a two-dimensional medium without flexing resistance which can only be subject to stretching in all directions.)

There is an important generalization of equations (60) and (62) known as the **Klein-Gordon equation**. It can be obtained if we

suppose that the string or the membrane is acted upon by a uniformly distributed linear restoring force directed toward the equilibrium position. Let the reader show that this reduces to adding a term of the form $-ku$ $(k>0)$ to the right-hand side of (60) or (62).

13. Vibrations of a Bar and of a Plate. Now we proceed to consider longitudinal vibrations of a rectilinear bar, whose axis occupies the position of the segment of the x-axis with end points $x=0$ and $x=l$, when the bar is in (unloaded) equilibrium state. Let $u(x, t)$ be the longitudinal displacement (along the x-axis) of the point x of the bar at time moment t. We shall suppose that the bar is under the action of a longitudinal external force distributed with density $f(x, t)$ along the bar and that there are elastic fixings at the ends of the bar, the elasticity factors being, respectively, α_0 and α_l. When the bar is deformed the relative elongation of its element dx is equal to $\frac{\partial u}{\partial x}$, and this creates an elastic force in the element. According to Hooke's law, this force is expressed by the formula

$$P=FE\frac{\partial u}{\partial x} \tag{63}$$

where F is the cross section of the bar and E is the **(longitudinal) modulus of elasticity (Young's modulus of tension)** characterizing the material of the bar. Therefore, the work performed by the internal forces when the bar passes from the state of equilibrium $u\equiv 0$ to the state described by the function $u(x, t)$ is equal to

$$dA=\frac{1}{2}FE\frac{\partial u}{\partial x}\partial_x u=\frac{1}{2}FEu_x'^2dx$$

(let the reader explain the appearance of the coefficient $\frac{1}{2}$). Consequently, the potential energy of the bar in an arbitrary state of strain is given by the formula

$$U=\int_0^l\left[\frac{1}{2}FEu_x'^2-f(x, t)u\right]dx+\frac{1}{2}\alpha_0u^2\Big|_{x=0}+\frac{1}{2}\alpha_lu^2\Big|_{x=l}$$

The kinetic energy of the bar is obviously equal to

$$T=\int_0^l\frac{1}{2}\rho Fu_t'^2dx$$

where ρ is the volume mass density of the material of the bar. Assuming that the bar is homogeneous ($\rho=\text{const}$) and applying

Hamilton's principle we arrive at the *differential equation of longitudinal vibrations of the bar*:

$$f(x, t) - \rho F u''_{tt} + F E u''_{xx} = 0$$

This equation coincides with equation (60) if we put $a = \sqrt{\frac{E}{\rho}}$ and $f_1 = \frac{f}{\rho F}$. According to Sec. 1.15 (see formula (1.84)) the natural boundary conditions are written in the form

$$(FEu'_x - \alpha_0 u)_{x=0} = 0, \quad (FEu'_x + \alpha_l u)_{x=l} = 0 \tag{64}$$

and are referred to as *(homogeneous) boundary conditions of the third kind*. (Let the reader derive conditions (64) directly from formula (63) and explain the meaning of the signs minus and plus in front of the quantities α_0 and α_l.) If one of the end points is free, for instance, the right one, we have $\alpha_l = 0$, and the corresponding boundary condition turns into

$$u'_x|_{x=l} = 0$$

The latter is called a *boundary condition of the second kind*. Finally, if there is a rigid fixing at the point $x = l$, we can (conditionally) put $\alpha_l = \infty$. Dividing the second condition (64) by α_l and making α_l tend to ∞ we receive a *boundary condition of the first kind*:

$$u|_{x=l} = 0$$

The next example treats of *transverse vibrations* of the same bar. We shall assume that in the process of vibrations the elements of the bar are only in the state of bending strain. When a rectilinear element dx of the bar is sudjected to bending it becomes curvilinear, its length takes on a value ds, and in the cross section of the bar there appears a restoring moment M. In the theory of small deformations it is assumed that M is proportional to the curvature $k = \left|\frac{d\varphi}{ds}\right|$ (e.g. see [107], Sec. VII.25) where $d\varphi$ is the angle of contingence of the arc ds, the proportionality factor μ (dependent on the properties of the material of the bar and on the moment of inertia of its cross section) being constant. The work that must be performed to give the element dx of the rectilinear bar this deformation is equal to $\frac{1}{2} M |d\varphi| = \frac{1}{2} \mu k^2 ds$. For the sake of generality, let as suppose that the bar is acted upon by a transverse external force distributed with density f. Now, taking into account that

$$ds = \sqrt{1 + u'^2_x}\, dx \quad \text{and} \quad k = |u''_{xx}| (1 + u'^2_x)^{-\frac{3}{2}}$$

and performing the linearization we arrive at the following expression for the action:

$$\int_{t_0}^{t_1} dt \int_0^l \left[\frac{1}{2}\rho u_t'^2 - \frac{1}{2}\mu u_{xx}''^2 + f(x, t)\, u\right] dx$$

The corresponding Euler-Ostrogradsky equation (the *equation of transverse vibrations of the bar*) is written as

$$\rho \frac{\partial^2 u}{\partial t^2} = -\mu \frac{\partial^4 u}{\partial x^4} + f(x, t)$$

The form of the boundary conditions is specified by the states of the ends of the bar. If an end is rigidly fixed the corresponding boundary condition is of the form

$$u = 0, \quad u_x' = 0 \tag{65}$$

(A more general variant of a rigid fixing appears when the end point is in a forced motion; then the right members of (65) are given functions of time.) If an end of the bar is supported the boundary condition is written as

$$u = 0, \quad u_{xx}'' = 0 \tag{66}$$

(formulas (66) follow from the equality $M = \mu k$ but can also be obtained as natural boundary conditions, which we leave to the reader). Now let us consider the case when there are (linear) elastic fixings at the end points of the bar providing resisting forces and moments opposing the bending and the displacements at these points. In this case we must add an expression of the form

$$\frac{1}{2}[\alpha_0 u^2 + \beta_0 u_x'^2]_{x=0} + \frac{1}{2}[\alpha_l u^2 + \beta_l u_x'^2]_{x=l}$$

to the potential energy and take the natural boundary conditions which assume the form

$$\left.\begin{aligned} (\mu u_{xx}'' + \beta_l u_x')_{x=l} = 0, \quad (\mu u_{xxx}''' - \alpha_l u)_{x=l} = 0 \\ (\mu u_{xx}'' - \beta_0 u_x')_{x=0} = 0, \quad (\mu u_{xxx}''' + \alpha_0 u)_{x=0} = 0 \end{aligned}\right\} \tag{67}$$

(check it up!). Boundary conditions (65) and (66) can be obtained from conditions (67) as limiting cases. Besides, if there are also external forces and external moments applied to the ends of the bar and performing some work when the ends are turned and displaced, we obtain nonhomogeneous conditions of form (67); let the reader analyse this case.

The process of vibrations of a *plate* which we are going to discuss now is a two-dimensional analogue of the problem considered above. We shall suppose that the unloaded homogeneous

plate is in the plane state of equilibrium. To obtain the expression of the potential energy of an element of the deformed plate located in the vicinity of a point M_0 let us choose a Cartesian coordinate system with origin at M_0 such that the xy-plane is tangent to the element under consideration. Then the equation of the deformed plate in the vicinity of the point M_0 is written in the form

$$z = \frac{1}{2}\left[(z''_{xx})_0 x^2 + 2(z''_{xy})_0 xy + (z''_{yy})_0 y^2\right] + \textit{terms of higher order of smallness} \tag{68}$$

By analogy with the one-dimensional case, we shall assume that, under small deformations, the potential energy of the element of the plate is a quadratic form in the partial derivatives of the second order of the function z with respect to x and y. The potential energy being independent of the choice of the coordinate axes, we can reduce the quadratic form in (68) to its principal axes and thus obtain the expression

$$dU = (Ak_1^2 + Bk_1k_2 + Ck_2^2)\,dS$$

of the potential energy of the element where k_1 and k_2 are the eigenvalues whose moduli are equal to the curvatures of the principal sections of the surface of the plate at the point M_0 (e.g. see [107], Sec. XII.9) and dS is the area of the element. We shall suppose that the plate is isotropic; then we have $A = C$. Denoting this common value of A and C by $\frac{\mu}{2}$ and using the relation between the eigenvalues as roots of the (quadratic) characteristic equation and the coefficients of the equation we obtain (check up this result!) the expression

$$dU = \frac{\mu}{2}(k_1^2 + k_2^2) + Bk_1k_2 = \frac{\mu}{2}(k_1 + k_2)^2 + (B - \mu)\,k_1k_2 =$$

$$= \frac{\mu}{2}(z''_{xx} + z''_{yy})^2 + (B - \mu)(z''_{xx}z''_{yy} - z''^2_{xy})$$

Hence, the potential energy of the plate in the state of strain when it is bent is expressed as

$$U = \frac{1}{2}\int\left[\mu(u''^2_{xx} + 2u''_{xx}u''_{yy} + u''^2_{yy}) + (B - \mu)(u''_{xx}u''_{yy} - u''^2_{xy})\right] dS$$

where $u = u(x, y, t)$ is the transverse deflection of the point (x, y) of the plate at moment t from its plane equilibrium state and x and y are the coordinates in this plane (which we again denote by the same letters x and y). When writing the latter integral we have passed from the projections of the elements of the plate on the corresponding tangent planes to their projections on the

plane $u=0$ corresponding to the state of equilibrium of the unloaded plate; the justification of this procedure lies in the fact that the discarded terms are of higher order of smallness in the case of small vibrations.

Now, forming the integral expressing the action and applying Hamilton's principle we deduce the *equation of vibrations of the plate*

$$\rho \frac{\partial^2 u}{\partial t^2} = -\mu \left(\frac{\partial^4 u}{\partial x^4} + 2 \frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} \right) \tag{69}$$

where $\rho = \text{const}$ is the areal mass density. The operator in the parentheses in expression (69) (applied to the function u) is referred to as the **biharmonic operator.** Using the symbol Δ for the Laplacian operator we can write this operator in the form

$$\frac{\partial^4}{\partial x^4} + 2 \frac{\partial^4}{\partial x^2 \partial y^2} + \frac{\partial^4}{\partial y^4} = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)^2 = \Delta^2 \tag{70}$$

and equation (69) (the **biharmonic equation**) in the form

$$\rho u_{tt}'' = -\mu \Delta^2 u$$

Let the reader consider the nonhomogeneous equation corresponding to (69) and also investigate the corresponding boundary conditions.

14. General Scheme for Applying Variational Principles to Physical Fields. As was mentioned in Sec. 12, the modified form of Hamilton's principle can be applied to a wide class of physical fields. In particular, in [105], Chapter 3, the reader can find many examples of this application (also see [48]). Here we shall confine ourselves to a number of general remarks concerning these problems and consider one typical example of a field.

Hamilton's principle in particle mechanics is based on the notion of the Lagrangian function; in applications of the principle to continua (continuous fields) the same role is played by the density of the Lagrange function. In mechanical problems this function is readily constructed because the expressions for the potential and kinetic energies can be easily found; in the case of a field of other nature this may not be so simple since it is sometimes difficult to distinguish between these types of energy. For instance, the energy of an electric charge moving in a magnetic field has much in common with kinetic energy (e.g. it depends on the velocity of motion) but at the same time resembles potential energy (because it depends on the state of the field as well). Therefore, we often apply artificial techniques similar to the one used in Sec. 10: we begin with considering some simpler problems and reducing the corresponding equations of motion to the form of Euler's equations, compare these prob-

lems and construct the Lagrangian function which is then generalized to more complicated problems. In the construction of the Lagrangian function a primary role is also played by scalar invariants of the physical problem we deal with because the Lagrangian function and its density are representable in terms of the invariants.

Now suppose that the density of the Lagrange function has been determined (in this section we shall use the latter L to designate the density). Usually the function L is expressed in terms of scalar variables $\psi_i = \psi_i(\xi_1, \xi_2, \xi_3, \xi_4)$, $i = 1, 2, \ldots, n$ (**field variables**), whose values characterize the state of the field at every given point (ξ_1, ξ_2, ξ_3) of the space at given time moment $t = \xi_4$, and in terms of their derivatives $\psi_{ik} = \frac{\partial \psi_i}{\partial \xi_k}$ $(i, k = 1, 2, \ldots, n)$. (Some of the field variables can be components of a vector or tensor.) Besides, in the case of a nonhomogeneous field or nonaffine coordinates the function L may also depend on ξ_1, ξ_2, ξ_3; for a nonautonomous system it can involve time $t = \xi_4$ as well. The 4-fold integral

$$\int\int\int\int L d\xi_1 d\xi_2 d\xi_3 d\xi_4 \tag{71}$$

which is a generalization of Hamilton's action integral must be invariant relative to the transformations permissible in the theory we deal with (e.g. in the theory of relativity it is invariant under Lorentz' transformations and the like). According to Sec. 1.11 the condition that the actual state of the system is characterized by a stationary value of integral (71) leads to the system of the Euler equations (the Euler-Ostrogradsky equations)

$$\sum_{k=1}^{4} \frac{\partial}{\partial \xi_k} \left(\frac{\partial L}{\partial \psi_{ik}} \right) - \frac{\partial L}{\partial \psi_i} = 0 \qquad (i = 1, 2, \ldots, n) \tag{72}$$

(in mechanics these equations are usually referred to as the **Lagrange equations** or the **Euler-Lagrange equations**).

If, in addition to quantities continuously distributed in space, we also consider masses, charges, etc. concentrated at points or distributed over lines and surfaces, expression (71) may include integrals of lower order taken over these lines and surfaces. These additional terms can be transformed to the form of 4-fold integrals and included into integral (71) if we use delta functions (e.g. cf. [107], Sec. XIV.25) although this procedure may prove inexpedient.

Furthermore, the addition of a *divergence expression* of the type

$$\sum_{k=1}^{4} \frac{\partial}{\partial \xi_k} A_k(\xi_1, \ldots, \xi_4, \psi_1, \ldots, \psi_n)$$

to the function L results in the appearance of an additional summand in expression (71) determined by the boundary conditions but does not affect the Euler-Lagrange equations (72) which remain invariant. For a more particular case this was indicated in Sec. 10. The latter property is sometimes used for simplifying the form of the density of the Lagrangian function.

The quantity

$$p_i=\frac{\partial L}{\partial \psi_{i4}} \qquad (i=1,\ 2,\ \ldots,\ n) \tag{73}$$

is the **canonical momentum density** (**generalized momentum density** in mechanics). Equations (72) can be written in the form

$$\frac{\partial p_i}{\partial t}=\frac{\partial L}{\partial \psi_i}-\sum_{k=1}^{3}\frac{\partial}{\partial \xi_k}\left(\frac{\partial L}{\partial \psi_{ik}}\right)$$

resembling Newton's equations of motion. The right-hand members of this system are analogues of the components of the density of a generalized force; the first term $\frac{\partial L}{\partial \psi_i}$ $(i=1,\ 2,\ \ldots,\ n)$ is usually connected with the presence of external forces applied to the field while the second term $-\sum_{k=1}^{3}\frac{\partial}{\partial \xi_k}\left(\frac{\partial L}{\partial \psi_{ik}}\right)$ is related to the action of the field upon the quantity ψ_i.

The square matrix $\mathbf{W}$ of the fourth order with components

$$W_{ij}=\sum_{r=1}^{n}\frac{\partial \psi_r}{\partial \xi_i}\frac{\partial L}{\partial \psi_{rj}}\ (i\neq j), \quad W_{ii}=\sum_{r=1}^{n}\frac{\partial \psi_r}{\partial \xi_i}\frac{\partial L}{\partial \psi_{ri}}-L \tag{74}$$

is termed the **stress-energy matrix.** The element

$$H=W_{44}=\sum_{r=1}^{n}\frac{\partial \psi_r}{\partial t}p_r-L=\sum_{r=1}^{n}\psi_{r4}p_r-L \tag{75}$$

of the matrix is the **energy density** of the field. If the function L, the spatial domain (G) occupied by the field and the values of ψ_i on the boundary of the domain (G) are independent of time t the integral $\iiint_{(G)} H\,d\xi_1\,d\xi_2\,d\xi_3$ does not depend on t either (check up this property).

Expressing the quantities ψ_{i4} (provided this is possible) in terms of the other variables from equalities (73) and substituting these expressions into (75) we obtain the representation of H as a function of the coordinates, field variables, their derivatives with respect to spatial coordinates and of the canonical momen-

tum density. Then rewriting integral (71) in the form

$$\iiiint \left(-H+\sum \frac{\partial \psi_r}{\partial t} p_r\right) d\xi_1\, d\xi_2\, d\xi_3\, d\xi_4$$

and forming the corresponding system of the Euler-Ostrogradsky equations (cf. Sec. 4) we arrive at the canonical form

$$\frac{\partial p_r}{\partial t}=-\frac{\partial H}{\partial \psi_r}+\sum_{k=1}^{3}\frac{\partial}{\partial \xi_k}\frac{\partial H}{\partial \psi_{rk}},\quad \frac{\partial \psi_r}{\partial t}=\frac{\partial H}{\partial p_r}\ (r=1,\ 2,\ \ldots,\ n) \qquad (76)$$

of system (72). (Let the reader prove that equations (72) and (76) are equivalent and derive, with the aid of a finite-dimensional approximation of the field constructed by analogy with the example considered in Sec. 1.11, the expressions for the right-hand sides of (76).)

Proceeding from expression (74), and using equations (72) we can easily verify that

$$\sum_{j=1}^{4}\frac{\partial W_{ij}}{\partial \xi_j}=-\frac{\partial L}{\partial \xi_i}\ (i=1,\ 2,\ 3,\ 4) \qquad (77)$$

(let the reader deduce this relation). When computing the left-hand sides of (77) which involve the total partial derivatives $\frac{\partial W_{ij}}{\partial \xi_j}$ we must take into account that W_{ij} depends on all ψ_r and ψ_{rs} while the derivatives $\frac{\partial L}{\partial \xi_i}$ on the right-hand sides are taken solely with respect to the variables ξ_i on which the function L depends explicitly.

Let us assume, for simplicity's sake, that ξ_1, ξ_2 and ξ_3 are Cartesian coordinates ($\xi_1=x$, $\xi_2=y$ and $\xi_3=z$) and put

$$\mathbf{S}=W_{41}\mathbf{i}+W_{42}\mathbf{j}+W_{43}\mathbf{k}$$

Then, if L does not depend on t, we obtain from (77), for $i=4$, the equation

$$\operatorname{div}\mathbf{S}+\frac{\partial H}{\partial t}=0$$

Consequently, the vector $\mathbf{S}$ is the **energy-flux density** of the field. Furthermore, the vector

$$\mathbf{P}=W_{14}\mathbf{i}+W_{24}\mathbf{j}+W_{34}\mathbf{k}$$

which is equal to the sum $\sum_{r=1}^{n} p_r \operatorname{grad}\psi_r$ (check it up!) is called the **momentum density of the field.** It follows from (77) that if

L is independent of ξ_1, ξ_2, ξ_3 we have

$$\sum_{i,\,j=1}^{3} \mathbf{e}_i \nabla \cdot (W_{ij}\mathbf{e}_j) = -\frac{\partial \mathbf{P}}{\partial t}$$

where $\mathbf{e}_1 = \mathbf{i}$, $\mathbf{e}_2 = \mathbf{j}$ and $\mathbf{e}_3 = \mathbf{k}$ are the Cartesian base vectors.

In conclusion we shall discuss the application of the Noether theorem (see Sec. 6) to continuous fields. For simplicity, assume that the field we deal with is defined in the whole space of the variables ξ_1, ξ_2, ξ_3 and is rapidly decreasing at infinity. Let integral (71) be invariant with respect to a one-parameter family of mappings which we shall write in the vector form

$$\bar{\xi} = \varphi(\xi, \psi, \psi'_\xi, \alpha), \quad \bar{\psi} = \Phi(\xi, \psi, \psi'_\xi, \alpha) \tag{78}$$

where α is a scalar parameter and ξ is the vector with components ξ_1, ξ_2, ξ_3, ξ_4, etc. The invariance is understood in the sense that if we take an arbitrary field $\psi(\xi)$ and compute, for an arbitrary value of α, the field variables $\bar{\psi}(\bar{\xi})$ expressed by formulas (78), we have the equality $\int L(\xi, \psi, \psi'_\xi)\, d\xi = \int L(\bar{\xi}, \bar{\psi}, \bar{\psi}'_\xi)\, d\bar{\xi}$ (where $d\xi = d\xi_1\, d\xi_2\, d\xi_3\, d\xi_4$ and the symbol $\int$ designates the 4-fold integral). According to Sec. 6, under these assumptions, the vector field

$$\left(\frac{\partial L}{\partial \psi_{ik}}\right)^* [\Phi_0 - (\psi_{ik})\,\varphi_0] + L\varphi_0 \text{ where } \varphi_0 = \frac{\partial \varphi}{\partial \alpha}\Big|_{\alpha=0}, \quad \Phi_0 = \frac{\partial \Phi}{\partial \alpha}\Big|_{\alpha=0}$$

has no sources of vector lines in the four-dimensional space of the vectors $\xi = (\xi_1, \xi_2, \xi_3, \xi_4)$. Then the flux of the field across any two hyperplanes $t = \xi_4 = \text{const}$ has one and the same value (why?), and thus we obtain the conservation law for the integral

$$\iiint \left\{\left(\frac{\partial L}{\partial \psi_{ik}}\right)^* [\Phi_0 - (\psi_{ik})\,\varphi_0] + L\varphi_0\right\}_t d\xi_1\, d\xi_2\, d\xi_3$$

whose value is retained, i.e. independent of time t (here the subscript t indicates the projection of the vector on the t-axis).

For example, suppose that the density L of the Lagrangian function is independent of one of the coordinates ξ_1, ξ_2, ξ_3, ξ_4, say of ξ_j. Then we can put $\bar{\xi}_j = \xi_j + \alpha$, $\bar{\xi}_k = \xi_k$ for $k \neq j$ and $\bar{\psi} = \psi$, whence $\varphi_0 = \mathbf{e}_j$ and $\Phi_0 = 0$. Thus, in this case the integral $\iiint W_{j4}\, d\xi_1\, d\xi_2\, d\xi_3$ is time-independent (let the reader check up this assertion). Consequently, if the density of the Lagrangian function is independent of the spatial coordinates we have the law of conservation of momentum and if it is independent of time, the law of conservation of total energy (the latter has already been proved).

In [48] the reader can find the discussion of the law of conservation of moment of momentum.

15. Equations of Motion of an Elastic Medium. Here we shall consider the application of the theory discussed in Sec. 14 to deriving the equations of motion of an isotropic homogeneous elastic medium. Let x_1, x_2, x_3 be Cartesian coordinates and $\mathbf{u}=\mathbf{u}(x_1, x_2, x_3, t)$ the field of displacements of the points of an elastic medium. In Sec. V.1.5 we introduced the notion of the strain tensor whose components are

$$\varepsilon_{ij}=\frac{1}{2}\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right) \tag{79}$$

In the theory of elasticity we also consider the (Cartesian) **stress tensor** with components σ_{ij} each of which is equal to the projection on the x_j-axis ($j=1, 2, 3$) of the force (per unit area) acting upon a small area whose outer normal is in the direction of the x_i-axis ($i=1, 2, 3$). Both tensors are symmetric, and in the case of small deformation (for which the linear law of elasticity holds) of an isotropic medium they are connected by the linear relation

$$\sigma_{ij}=\lambda\varepsilon_{rr}\delta_{ij}+2\mu\varepsilon_{ij}$$

(understood in the sense of the tensor notation; see Sec. V.1.1) in which the constants λ and μ are specified by the elasticity properties of the medium and are connected with other characteristics of the medium. For instance, in the case of tension along the x_1-axis (i.e. when $\sigma_{ij}=Q\delta_{i1}\delta_{j1}$) we have (check it up!)

$$\varepsilon_{11}=\frac{\lambda+\mu}{\mu(3\lambda+2\mu)}Q,\quad \varepsilon_{22}=\varepsilon_{33}=-\frac{\lambda}{2\mu(3\lambda+2\mu)}Q,\quad \varepsilon_{12}=\varepsilon_{23}=\varepsilon_{13}=0$$

Therefore, the quantity $\frac{\mu(3\lambda+2\mu)}{\lambda+\mu}$ is nothing but the modulus of elasticity entering into formula (63). The quantity $\frac{\lambda}{2(\lambda+\mu)}$ which is the ratio of the transverse contraction to the longitudinal extension is known as the **Poisson ratio.** In the case of uniform compression (that is when $\sigma_{ij}=-Q\delta_{ij}$) we obtain the relation $\varepsilon_{ij}=\frac{-Q\delta_{ij}}{3\lambda+2\mu}$. But the sum $\sum_{i=1}^{3}\varepsilon_{ii}$ is equal to the *volumetric expansion coefficient*, and thus the constant $\frac{3\lambda+2\mu}{3}$ is the **bulk** (or **compressibility**) **modulus** of the isotropic elastic medium.

Let us isolate (mentally) an elementary cube with edges parallel to the coordinate axes and consider its deformation corresponding to the passage from the equilibrium state (in which it is unloaded) to a given state of strain, as the stresses vary from zero to some given values. Then the density of the potential

energy of this state of strain is expressed by the formula

$$U=\frac{1}{2}\sigma_{ij}\varepsilon_{ij}=\frac{\lambda}{2}(\varepsilon_{rr})^2+\mu\varepsilon_{ij}\varepsilon_{ij}$$

(whose derivation is left to the reader). By (79), it follows that

$$U=\frac{\lambda}{2}(\operatorname{div}\mathbf{u})^2+\mu\left[\left(\frac{\partial u_1}{\partial x_1}\right)^2+\left(\frac{\partial u_2}{\partial x_2}\right)^2+\left(\frac{\partial u_3}{\partial x_3}\right)^2\right]+$$
$$+\frac{\mu}{2}\left[\left(\frac{\partial u_1}{\partial x_2}+\frac{\partial u_2}{\partial x_1}\right)^2+\left(\frac{\partial u_1}{\partial x_3}+\frac{\partial u_3}{\partial x_1}\right)^2+\left(\frac{\partial u_2}{\partial x_3}+\frac{\partial u_3}{\partial x_2}\right)^2\right]$$

The density of the kinetic energy is obviously given by the formula

$$T=\frac{\rho}{2}\left|\frac{\partial\mathbf{u}}{\partial t}\right|^2=\frac{\rho}{2}\left[\left(\frac{\partial u_1}{\partial t}\right)^2+\left(\frac{\partial u_2}{\partial t}\right)^2+\left(\frac{\partial u_3}{\partial t}\right)^2\right]$$

where $\rho=\text{const}$ is the mass density of the medium. Consequently, we can write the expression of the density $L=T-U$ of the Lagrangian function in which the role of the field variables ψ_i is played by the three projections u_i $(i=1, 2, 3)$ of the displacement vector. The Euler-Lagrange equations of form (72) corresponding to this function L are written (check it up) in the form

$$\rho\frac{\partial^2 u_i}{\partial t^2}-\lambda\frac{\partial}{\partial x_i}\operatorname{div}\mathbf{u}-\mu\nabla^2 u_i-\mu\frac{\partial}{\partial x_i}\operatorname{div}\mathbf{u}=0 \qquad (i=1, 2, 3)$$

Multiplying these equations by the unit vectors of the corresponding coordinate axes and adding together the results we arrive at the (vector) *equation of free oscillations of the isotropic homogeneous elastic medium*:

$$\rho\mathbf{u}''_{tt}=(\lambda+\mu)\ \text{grad div}\ \mathbf{u}+\mu\nabla^2\mathbf{u}$$

Let the reader show that the density H of the total energy of the field is equal to $T+U$, the energy-flux density $\mathbf{S}$ is equal to $-\sigma\mathbf{u}'_t$ where $\sigma=(\sigma_{ij})$ is the matrix of the stress tensor and the momentum density $\mathbf{P}$ is equal to $\rho(\nabla\mathbf{u})\cdot\mathbf{u}'_t$ where the expression in the parentheses is the (formal) tensor product of the symbolic vector ∇ by the vector $\mathbf{u}$ (see Sec. V.1.3). Also verify that the "spatial part" $\mathbf{W}'$ of the matrix $\mathbf{W}$ (formed of the elements belonging to the first three rows and columns of $\mathbf{W}$) is expressed by the tensor formula $\mathbf{W}'=-(\nabla\mathbf{u})\cdot\sigma-L\mathbf{I}$.

16. Dissipative Systems. Here we shall confine ourselves to autonomous dissipative systems. As was indicated in Sec. 11, an autonomous system to which Hamilton's principle is applicable must be conservative while the total energy of a dissipative system decreases. But it turns out that there is an artificial technique which makes it possible to extend Hamilton's principle to linear nonconservative systems and, in particular, to systems in which

the energy is dissipated (for instance, due to friction). To this end, we (mentally) add to the original dissipative system another system which is its "opposite" in the sense that it possesses the same properties and parameters but absorbs the energy lost by the original system (thus, the second system is one with *negative friction*). Then the total system composed of the original, real, system and the imagined system is conservative and admits of the application of Hamilton's principle.

Let us consider the collection of the coordinates $q_1, q_2, \ldots, q_n$ determining the state of the original system as the column number vector $\mathbf{q}$ and suppose that the system of differential equations of motion is written in the vector form

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{R}\dot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{0} \tag{80}$$

where $\mathbf{M}$, $\mathbf{R}$ and $\mathbf{K}$ are constant square matrices of order n. (In the simplest case of a system of linear oscillators with elastic connections the matrix $\mathbf{M}$ is diagonal.) Then the "opposite" system absorbing the energy lost by the original system (equation (80)) can be described by the vector equation

$$\mathbf{M}^*\ddot{\mathbf{s}} - \mathbf{R}^*\dot{\mathbf{s}} + \mathbf{K}^*\mathbf{s} = 0 \tag{81}$$

where $\mathbf{s}$ is the number vector composed of the coordinates s_1, $s_2, \ldots, s_n$ of this system and the asterisk designates the transposed vectors and matrices. The physical meaning of the plus sign in front of the second term is that the positive friction (described by the term $\mathbf{R}\dot{\mathbf{q}}$ in equation (80)) is replaced by the equivalent negative friction.

It is readily seen (check it up!) that equations (80) and (81) considered together constitute the system of Euler's equations corresponding to the functional

$$\int_{t_0}^{t_1} \left(\dot{\mathbf{s}}^*\mathbf{M}\dot{\mathbf{q}} + \frac{1}{2}\dot{\mathbf{s}}^*\mathbf{R}\mathbf{q} - \frac{1}{2}\mathbf{s}^*\mathbf{R}\dot{\mathbf{q}} - \mathbf{s}^*\mathbf{K}\mathbf{q} \right) dt$$

In proving this property it is advisable to use the fact that if $\frac{df}{d\mathbf{x}}$ $(\mathbf{x} = (x_1, \ldots, x_m))$ is understood as the column number vector with components $\frac{\partial f}{\partial x_i}$, we have the relation

$$\frac{d}{d\mathbf{x}}(\mathbf{a}^*\mathbf{x}) = \frac{d}{d\mathbf{x}}(\mathbf{x}^*\mathbf{a}) = \mathbf{a}$$

Thus, the function

$$L = \dot{\mathbf{s}}^*\mathbf{M}\dot{\mathbf{q}} + \frac{1}{2}\dot{\mathbf{s}}^*\mathbf{R}\mathbf{q} - \frac{1}{2}\mathbf{s}^*\mathbf{R}\dot{\mathbf{q}} - \mathbf{s}^*\mathbf{K}\mathbf{q}$$

can be taken as the Lagrangian function for this combined system (in contrast to autonomous systems considered before this function contains linear terms with respect to derivatives but this does not affect the validity of our considerations). Computing momenta p_i corresponding to the variables q_i and momenta r_i associated with the variables s_i and forming the Hamiltonian function we obtain (check it up!)

$$\mathbf{p}=\mathbf{M}^*\dot{\mathbf{s}}-\frac{1}{2}\mathbf{R}^*\mathbf{s},\quad \mathbf{r}=\mathbf{M}\dot{\mathbf{q}}+\frac{1}{2}\mathbf{R}\mathbf{q},$$

$$H=\dot{\mathbf{s}}^*\mathbf{M}\dot{\mathbf{q}}+\mathbf{s}^*\mathbf{K}\mathbf{q}=\left(\mathbf{p}^*+\frac{1}{2}\mathbf{s}^*\mathbf{R}\right)\mathbf{M}^{-1}\left(\mathbf{r}-\frac{1}{2}\mathbf{R}\mathbf{q}\right)+\mathbf{s}^*\mathbf{K}\mathbf{q}$$

(of course, the matrix $\mathbf{M}$ is supposed to be nonsingular). By virtue of Sec. 2, the function H thus constructed assumes a constant value for any pair of solutions of equations (80) and (81).

17. Principle of Minimum of the Potential Energy. In this section we shall only deal with conservative autonomous systems. Let us consider a system with finite number n of degrees of freedom whose generalized coordinates are $q_1, q_2, \ldots, q_n$. Suppose that the system is in an equilibrium state $q^0=(q_1^0, q_2^0, \ldots, q_n^0)$. Since this state can be regarded as a particular case of motion we can apply Hamilton's principle (Sec. 8) to it and form Euler's equations which are implied by the principle. For the state of equilibrium Euler's equations reduce to the simple equalities

$$U'_{q_i}(q_1^0, q_2^0, \ldots, q_n^0)=0 \qquad (i=1, 2, \ldots, n)$$

Thus, *every state of equilibrium of the system is a stationary point of the potential energy.*

Here we are going to discuss the stability of the state of equilibrium which, as will be shown, essentially depends on the type of the stationary value of the potential energy, that is on whether it is maximum, minimum or minimax value. Let the function $U(q)$ have an isolated local minimum at the point q^0 (see Fig. 106 illustrating this situation for $n=2$). Then it can easily be shown that *this point specifies a (Lyapunov) stable equilibrium state* (on the stability in the sense of Lyapunov see, for instance, [107], Sec. XV.22), which means that if the coordinates q_i, corresponding to the equilibrium state, are given arbitrary infinitesimal variations at an initial time moment t_0 and the system is given arbitrary infinitesimal velocities $\dot{q}_i$, the system will remain infinitely close to the state q^0 for all $t>t_0$. Indeed, take an arbitrarily small fixed neighbourhood (G) enclosing the point q^0, the boundary of (G) being denoted by (Γ) (see Fig. 106). Then the least value $U_{(\Gamma)}$ of the function U on (Γ) exceeds $U_0=U(q^0)$ (why?). On the other hand, the total energy H

of the system will be less than $U_{(\Gamma)}$ for any sufficiently small variations of the initial values of q and $\dot{q}$ (at the moment t_0) which were mentioned above. The total energy being time-independent (see Sec. 8), the system can never reach a point belonging to (Γ) (why?). Hence, it always remains in the neighbourhood (G), which is what we set out to prove.

By the way, this proof not only applies to conservative systems but to dissipative ones as well (Sec. 11).

On the other hand, it can be proved that if the function $U(q)$ has a maximum or a minimax at the point $q=q^0$, the equilibrium state $q=q^0$ is unstable. This property is visualized if we take, as a system with potential energy $U(q)$, a material point of mass m which is in free motion in the surface (shown in Fig. 106) whose equation is $U=U(q)$ under the action of the field of gravitation of intensity $g=\frac{1}{m}$ directed vertically downward. An intermediate case appears when the function U has a nonisolated minimum at the point q^0 (for instance the function $U(q_1, q_2)=q_1^2$ has the line of minimum $q_1=0$). The above visual interpretation (which can be confirmed by more rigorous considerations) indicates that in this case the conservative system is unstable (but this does not apply to dissipative systems!). In fact, if U has a nonstrict minimum, the system may move away by a finite distance from the original position when it is given arbitrarily small initial velocities although, in contrast to the case of a maximum or a minimax, the velocities remain small and do not attain finite values.

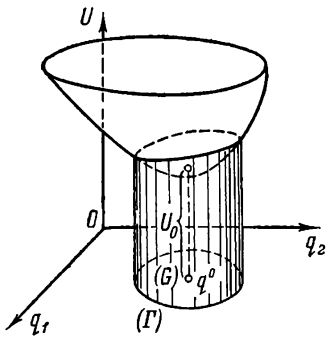


Fig. 106

Thus, the problem of determining the equilibrium states of a system with finite number n of degrees of freedom and testing them for stability reduces to the problem of finding and investigating the stationary points of a function of n arguments. Therefore, the analogous problem for continuous media (for which $n=\infty$) can be solved with the aid of variational methods, and if the stability is not implied by the formulation of the problem (for instance, by some physical considerations), we should resort to sufficient conditions for minimum when testing the stability.

18. Examples.

1. Let us consider the problem of determining the form of equilibrium of a uniform absolutely flexible inextensible thread suspended from two points in the homogeneous field of gravitation. Take a Cartesian coordinate system in which the x-axis is

horizontal and the y-axis is directed vertically upward. Assume, for simplicity, that the ends of the thread are at the points $(-b, 0)$ and $(b, 0)$, its shape being described by a function $y=y(x)$. According to Sec. 10, the sought-for function $y(x)$ must minimize the functional

$$\int yg\,dm = g\rho \int_{-b}^{b} y\sqrt{1+y'^2}\,dx$$

(where ρ is the linear density of the thread and g the acceleration of gravity) under the integral constraint

$$\int_{-b}^{b} \sqrt{1+y'^2}\,dx = L = \text{const} \tag{82}$$

(where $L > 2b$ is given) for the boundary conditions

$$y(-b)=0, \quad y(b)=0 \tag{83}$$

The standard scheme described in Sec. 1.12 leads to the general solution

$$y = C_1 \cosh\frac{x-C_2}{C_1} + \lambda$$

of the Euler equation where C_1 and C_2 are arbitrary constants and λ the Lagrange multiplier. Determining C_1, C_2 and λ from boundary conditions (83) and the equation of the constraint (relation (82)) we find the required particular solution $y = C_1\left(\cosh\frac{x}{C_1} - \cosh\frac{b}{C_1}\right)$ where

$$C_1 \sinh\frac{b}{C_1} = L \tag{84}$$

Thus, the solution is a catenary (see Sec. 1.8), the problem we have solved here being spoken of as the **problem of the catenary**. The left-hand side of (84) is a monotone decreasing function of C_1 which varies from ∞ to b as C_1 runs from 0 to ∞ (check it up!), and consequently there is exactly one state of equilibrium. Although it is obvious that this form of the thread is stable the reader can test it for stability with the aid of the method of Sec. 2.5 by showing that the function $y(x)$ found above gives the functional the required conditional minimum.

2. Let us investigate the stability of a rectilinear elastic bar subjected to **buckling** by a central force. Suppose that the axis of the unloaded bar goes along the x-axis, its ends being at the points $x=0$ and $x=l$. If a concentrated compressive force P parallel to the x-axis is applied to the free end $x=l$ of the bar while the end $x=0$ is rigidly fixed, the axis of the bar becomes

curvilinear. Let s be the arc length of the axis of the bar reckoned from the point $s=x=0$, and $\varphi=\varphi(s)$ be the angle of inclination to the x-axis of the element of the deformed bar with end points s and $s+ds$. Then, according to Sec. 13, we obtain the expression

$$U=\int_0^l \frac{1}{2}\mu k^2\,ds+P\Delta x\Big|_{s=l}=\int_0^l\left(\frac{1}{2}\mu\varphi_s'^2+P\cos\varphi-P\right)ds \qquad (85)$$

for the potential energy of the deformed bar. As is seen, it is sufficient to consider the function $\varphi(s)$ without introducing Cartesian coordinates. Functional (85) is taken with the boundary conditions

$$\varphi|_{s=0}=0,\quad \varphi_s'|_{s=l}=0 \qquad (86)$$

The Euler equation has the form

$$-\mu\varphi_{ss}''-P\sin\varphi=0 \qquad (87)$$

and its general solution can be expressed in terms of elliptic functions (Sec. II.3.14). But here we shall not dwell on this question because we only are interested in whether the extremal $\varphi\equiv 0$ gives a minimum to integral (85) under conditions (86). The results of Sec. 2.10 indicate that to answer this question we should determine the sign of the least eigenvalue $\lambda_{\min}$ of the linearized equation corresponding to (87). Thus we are interested in the least value of λ for which the boundary-value problem

$$-\mu\varphi''-P\varphi=\lambda\varphi,\ \varphi|_{s=0}=0,\ \varphi'|_{s=l}=0 \qquad (88)$$

possesses a nontrivial solution. The direct integration of the equation (which we leave to the reader) shows that

$$\lambda_{\min}=\frac{\pi^2}{4l^2}\mu-P$$

Thus, the *critical force* is

$$P_{cr}=\frac{\pi^2\mu}{4l^2}$$

If $P<P_{cr}$ the rectilinear state of equilibrium of the bar is stable, and if $P>P_{cr}$ it is unstable.

Usually it is more convenient to use dimensionless combinations of physical quantities because, for instance, given a compressive force P, we cannot immediately say whether the bar is stable or what is the stability factor and the like. In the above problem a combination of that kind is the quantity ("dimensionless force")

$$p=\frac{Pl^2}{\mu} \qquad (89)$$

which is proportional to the actual force P but also includes the parameters l and μ. Its critical value is equal to the (dimensionless) number $\frac{\pi^2}{4}=2.467$.

In more complicated problems of this type we have to resort to approximate calculations when determining the value of λ_{min}. If a problem of that kind includes a parameter α it often occurs that the stability is indicated, for a value $\alpha=\alpha_0$ of the parameter, by physical properties or some other considerations (see the beginning of Sec. 2.6). If we compute the values of $\lambda_{min}(\alpha)$ beginning with $\alpha=\alpha_0$, the first point α at which λ_{min} changes its sign specifies α_{cr}. If there are two parameters α and β we can similarly find the value of α_{cr} for a fixed β and then, varying β, construct the stability regions of the system in the $\alpha\beta$-plane.

In a problem involving a parameter α it is sometimes more convenient to apply the following argument. Since an eigenvalue is usually continuous in α, it must turn into zero when changing the sign. Conversely, if a continuous function $\lambda(\alpha)$ vanishes for some $\alpha=\tilde{\alpha}$, it usually changes its sign at the point $\tilde{\alpha}$ (unless the additional equality $\lambda'(\tilde{\alpha})=0$ holds). Therefore, to determine α_{cr} we can put $\lambda=0$ in the boundary-value problem and find the corresponding values of α among which the one nearest to α_0 is equal to α_{cr}. For example, passing to the dimensionless length $\sigma=\frac{s}{l}$ in problem (88) and putting $\lambda=0$, we obtain

$$\frac{d^2\varphi}{d\sigma^2}+p\varphi=0, \quad \varphi|_{\sigma=0}=0, \quad \left.\frac{d\varphi}{d\sigma}\right|_{\sigma=1}=0$$

(see (89)) whence we find the possible values $p=\frac{\pi^2}{4}, \frac{9\pi^2}{4}, \ldots$. For small $p>0$ the solution is stable, and therefore $p_{cr}=\frac{\pi^2}{4}$. For $\frac{\pi^2}{4}<p<\frac{9\pi^2}{4}$ we have one negative eigenvalue, for $\frac{9\pi^2}{4}<p<\frac{16\pi^2}{4}$ two negative eigenvalues, etc. More complicated problems may involve eigenvalues which are not monotone functions of the parameter, and then there can be several intervals of stability.

19. Barrier Potential. Let us consider a conservative autonomous system with one degree of freedom whose potential energy $U=U(q)$ (where q is the coordinate of the system) has the graph shown in Fig. 107. Here we have the six possible states of equilibrium $q^1, q^2, \ldots, q^6$ among which the three states q^1, q^3 and q^5 are stable. Let us consider one of these stable states of the system, for instance, q^3. The least of the differences $U(q^2)-U(q^3)$ and $U(q^4)-U(q^3)$ whose value we denote by K^3 (in Fig. 107 the first difference is shown as the least) is the **barrier potential** corresponding to the region shaded in Fig. 107, the latter being the

potential pit enclosing the equilibrium state q^3. The meaning of this terminology is that if the system is in the state q^3 and then is given an amount of kinetic energy less than K^3, the system cannot leave the shaded region, the pit (Fig. 107), because the total energy retains its value while the kinetic energy is nonnegative. Besides, if there is an arbitrarily small dissipation of energy, then, in a sufficiently long time, the system will return to the original state. We also have the same situation when the kinetic energy is increased any number of times provided that its total amount is less than the barrier potential. The notions of the potential pit

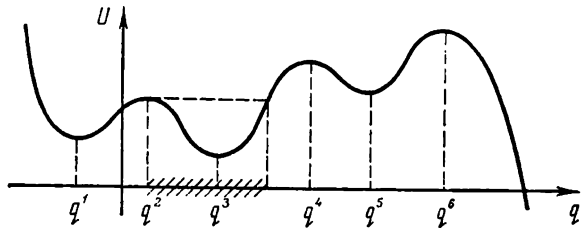


Fig. 107

and the barrier potential are similarly introduced for the other stable states of equilibrium of the system. The region $q > q^6$ in Fig. 107 can also be regarded as a potential pit: when the point representing the state of the system leaves this region the system recedes to infinity, i.e. $q \to \infty$.

Besides the points of minimum of the potential energy $U(q)$ $(q = (q_1, q_2, \ldots, q_n))$, a system with finite number n of degrees of freedom may possess manifolds in the space $q_1, q_2, \ldots, q_n$ on which U is equal to constant minimum values ("manifolds of minimum"). In a manifold of this type the system is in a **state of neutral equilibrium.** Usually these manifolds appear when the system possesses certain symmetry properties. The compact manifolds of neutral equilibrium are investigated in the same way as the isolated points of minimum of the potential energy, and therefore in what follows we shall apply the term a "point of minimum" both to isolated points of minimum and to manifold of minimum. If q^0 is one of the points of minimum of $U(q)$, then, to determine the corresponding barrier potential, we must join the point q^0 with all the other points of minimum and with infinity by all the possible paths (l), find the greatest value $U_{(l)}$ on each of the paths and choose the line (l) for which the value of $U_{(l)}$ is the least. The line (l) on which the quantity $U_{(l)}$ attains its minimum value is the path for which the system should be given the least amount of kinetic energy causing it to pass to another equilibrium state or to travel toward infinity (let the

reader carefully examine Fig. 107 which illustrates all these possibilities). After the path (l) has been determined we obtain the barrier potential corresponding to the state of equilibrium q^0; it is equal to the difference

$$K^0 = \min_{(l)} \max_{M \in (l)} U(M) - U(q^0) \tag{90}$$

The determination of quantity (90) reduces to a variational problem because the expression $U_{(l)} = \max_{M \in (l)} U(M)$ is a functional defined for the lines (l) belonging to the class described above, and thus we deal with a minimum problem for this functional. But the classical method of the calculus of variation when applied to this problem usually yields only a trivial result: the condition that the variation of the functional is equal to zero shows that at the point where the greatest value on the line (l) (giving minimum (90) to the functional) is achieved the function U has an unconditional stationary value (as a rule this is a point of minimax or a point belonging to a minimax manifold; there can also be a singular case when this point belongs to an $(n-1)$-dimensional manifold of maximum). But outside a small neighbourhood of that point there is a great arbitrariness in the possible structure of the line on which the minimum is reached, and therefore we cannot speak about a differential equation of the type of Euler's equation for this line. *Nonclassical minimax problems* of this kind have been widely investigated in recent years (in particular, see Sec. IV.5.7) although they were sometimes dealt with before.

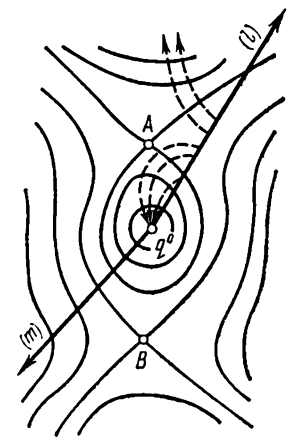


Fig. 108

The potential pit corresponding to the point q^0 can be determined as the connected component, containing the point q^0 (in the space q which is the **configuration space of the system**) of the set of all the states of the system satisfying the condition $U(q) < U(q^0) + K^0$. In the case $n = 2$ we can visualize this rule in the following manner. Suppose that there is a source at the point $(q_1^0, q_2^0, U(q_1^0, q_2^0))$ which fills with water the hollow whose bottom has the equation $U = U(q_1, q_2)$; then the moment the water overflows, the depth of the "lake" is equal to K^0 and its water-table covers the potential pit corresponding to the state q^0 (see Fig. 107).

The determination of the minimax on the boundary of a potential pit provides a numerical method for computing the corresponding barrier potential. For simplicity, let $n=2$ and let the level lines of the surface $U=U(q_1, q_2)$ have the shape shown in Fig. 108 (of course, in a real problem these level lines may be unknown). Suppose that we know an approximate form of the path (l) along which the system leaves the pit when it is given an amount of kinetic energy exceeding the barrier potential. If we take a sequence of points on the line (l) and move from each point in the direction of steepest descent along the trajectories determined by the equation $\dot{q}=-\text{grad } U(q)$, the trajectories starting near the point q^0 return to q^0 while those starting far from q^0 lead to another point of minimum or to infinity. Therefore, the presence of two neighbouring trajectories of steepest descent which diverge, in the above sense, indicates that they start in the vicinity of the sought-for minimax whose location can then be determined more precisely with the aid of Newton's method (e.g. see [107], Sec. XII.12) applied to the system of equations $U'_{q_i}=0$ $(i=1, 2, \ldots, n)$. (It should be noted that there can be points of minimax lying inside the potential pit, and therefore it is necessary that the trajectories of steepest descent we have spoken about not only diverge but lead to different minima.) If there are several paths along which the system can leave the pit (see the paths (l) and (m) in Fig. 108) the barrier potential is equal to the least height of the "passes". If there is no information about the paths along which the system can leave the pit we can test several paths taken at random and then choose the one with the lowest "pass".

The barrier potential of a system with infinite number of degrees of freedom (a continuous medium) can be determined on the basis of a finite-dimensional approximation.

20. Variational Principle for Conformal Mappings. Here we shall consider a variational principle, differing from those studied before, in which analytic functions are compared on the basis of a certain criterion. This principle can be used for an approximate construction of a conformal mapping (see § II.1) of one domain onto another.

Let us begin with an analytic mapping

$$f(z)=z+c_2z^2+c_3z^3+\ldots \tag{91}$$

of the circle (K_R): $|z|<R$ on a domain (H). According to Sec. II.1.4, the element of area dK undergoes the $|f'(z)|^2$-fold uniform magnification under the mapping, and therefore the area of the

domain (H) is expressed by the formula

$$H = \int\limits_{(K_R)} |f'(z)|^2 \, dK = \iint\limits_{(K_R)} f'(z)\,(f'(z))^* \rho \, d\rho \, d\varphi \qquad (z = \rho e^{i\varphi})$$

(if the mapping is multivalent the area is counted with the corresponding multiplicity). Substituting expansion (91) into the last integral, opening the brackets and integrating first with respect to φ and then with respect to ρ we obtain (check it up!)

$$H = \pi R^2 (1 + 2|c_2|^2 R^2 + 3|c_3|^2 R^4 + \ldots)$$

Thus, *under mapping* (91) *the area of the circle* (K_R) *always increases* (except the trivial case of the identity mapping which, of course, preserves the area).

Now we proceed to investigate the problem of constructing a conformal mapping $w = \varphi_0(z)$ of a given 1-connected domain (G) containing the origin $z = 0$ onto a circle $|w| < R$ (whose radius R is not specified beforehand) under the additional normalization condition

$$\varphi_0(0) = 0, \quad \varphi_0'(0) = 1 \tag{92}$$

The existence of a mapping of this type is implied by the Riemann theorem (Sec. II.1.16). Note that here we are not only given the angle of rotation of an infinitesimal neighbourhood of the point $z = 0$ but its magnification coefficient as well; the arbitrariness of the radius of the circle which is not known beforehand compensates for these exhausted degrees of freedom. It can readily be shown that the area of that circle is less than the area of any other region onto which the domain (G) can be mapped by means of an analytic function satisfying conditions (92), and consequently the function $\varphi_0(z)$ is the solution of the problem of determining the absolute minimum of the integral

$$\int\limits_{(G)} \varphi'(z)\,(\varphi'(z))^* \, dG \qquad (z = x + iy, \; dG = dx\,dy) \tag{93}$$

attained on the class of all analytic functions in (G) satisfying conditions (92). To prove this property it is sufficient to apply the above assertion concerning the area of the circle K_R to the function $\varphi(\varphi_0^-(w))$ where φ_0^- is the inverse function of φ_0 (let the reader carefully examine this argument).

In [58], Chapter V, there is given a technique for the application of this principle to constructing an approximate conformal mapping of a given region on a circle which is a modification of the *Ritz method* (see Sec. 4.1): the function $\varphi'(z)$ is represented approximately in the form $\sum\limits_{k=1}^{n} c_k g_k(z)$ where $g_k(z)$ are

certain analytic functions (for instance, $g_k(z) = z^{k-1}$), the constant parameters c_k being chosen in such a way that integral (93) is minimized under the condition $\sum_k c_k g_k(0) = 1$. In [58] the reader can find another variational principle which can also be used for solving the above problem: among all analytic functions in (G) satisfying conditions (92), the function $\varphi_0(z)$ is the one giving the absolute minimum to the integral

$$\int_{(L)} |f'(z)|\, dL$$

(where (L) is the contour bounding the domain (G)) expressing the length of the conformal image of the contour (L).

§ 4. DIRECT METHODS

The classical method for solving variational problems based on their reduction to Euler's differential equations often proves ineffective despite its great theoretical significance. Although we have studied a number of problems for which Euler's equations are integrable by quadratures and their solutions can even be expressed in terms of elementary functions, more complicated problems, particularly those involving functions of several variables, do not possess such simple solutions. Therefore, as a rule, we have to resort to a numerical method of solving the corresponding linear or nonlinear boundary-value problem for Euler's differential equation. Modern mathematics provides some techniques for solving such problems which are rather complicated and shall not be treated in our course. There is another group of numerical methods developed for solving variational problems in their original form without passing to Euler's equations which are referred to as **direct methods in variational problems.** In this section we shall discuss some of them.

Some information on direct methods and examples of their application can be found in the books indicated at the beginning of the present chapter. A thorough theoretical investigation of these methods and many examples are given in [58], Chapter IV, in [100] and [102].

Roughly speaking, there are two basic groups of approximate methods used in mathematical analysis for determining unknown functions: one of them is based on the reduction of the number of degrees of freedom while the other involves descrete schemes. (There are, of course, methods of "mixed" type and also those which do not belong to these groups either.) In the methods of the first group the independent variables are varied continuously but the unknown function is looked for in a special form involving some parame-

ters which are determined in such a way that the conditions of the problem are satisfied in the best manner. For instance, the interpolation method and the method of least squares are typical examples of that kind. In the application of these methods to complicated problems a primary role is played by physical intuition and analytic techniques because the effectiveness of the method essentially depends on the proper choice of the form of the sought-for solution involving a small number of parameters and on the criterion for estimating the approximations. In the methods of the second group (referred to as **net methods**) the unknown function is at the very beginning replaced by a collection of its values at the nodes of a net; typical examples of that kind are the methods of numerical integration of ordinary differential equations. Net methods are usually more universal and admit of elaboration of general algorithms, which makes them extremely useful for solving problems on electronic computers.

1. The Ritz Method for a Quadratic Functional. The Ritz method inaugurated in 1908 by W. Ritz, a German physicist and mathematician, is a typical method based on the reduction of the number of degrees of freedom. To begin with, we shall apply the Ritz method to the simpler case of functional (1.18) with boundary conditions (1.26) under the assumption that the functional is quadratic:

$$I\{y\} = \int_a^b [P(x)\, y'^2 + Q(x)\, y^2 + R(x)\, y]\, dx \tag{1}$$

Note that here we take a more general (nonhomogeneous) functional than in Sec. II.4. We can also include the terms with yy' and y' but this does not lead to a greater generality since these terms can be eliminated by means of integration by parts although in concrete cases the problem can be solved without this simplification.

The idea of the **Ritz method** is that the approximate expression of the extremal is constructed in the form

$$y = g_0(x) + \sum_{k=1}^{n} c_k g_k(x) \tag{2}$$

where $g_0(x)$, $g_1(x)$, ..., $g_n(x)$ are certain functions (**coordinate functions**) which are chosen in an appropriate manner, while the parameters c_k are determined on the basis of the condition that the value of functional (1) must be stationary. It is preferable to choose the function $g_0(x)$ in such a way that it satisfies conditions (1.26) and is as close as possible to the sought-for solution (of course, the latter requirement can only be met if we

have some information about the properties of the solution). (By the way, after the function $g_0(x)$ has been chosen we sometimes make the substitution $y = g_0(x) + z$ in order to obtain the homogeneous boundary conditions $z(a) = z(b) = 0$.) The coordinate functions $g_k(x)$ $(k = 1, 2, \ldots, n)$ are chosen so that they satisfy homogeneous conditions of type (1.26), that is $g_k(a) = g_k(b) = 0$. Then the fulfilment of conditions (1.26) for sum (2) is guaranteed for any values of c_k.

In theoretical studies the collection of the functions $g_1, \ldots, g_n$ is regarded as a part of an infinite system of functions

$$g_1(x),\ g_2(x),\ \ldots,\ g_n(x),\ g_{n+1}(x), \ldots \tag{3}$$

which is *linearly independent* and *complete* in the space of the functions $f \in C_1[a, b]$ satisfying the boundary conditions $f(a) = f(b) = 0$. The linear independence is understood in the sense that neither of the functions $g_j(x)$ is a linear combination of a finite number of other functions, and the completeness (see Sec. 1.3) means that any functions $f(x)$ can be approximately represented with an arbitrary accuracy (relative to the metric of the space C_1) as a linear combination of a finite number of functions (3). If the structure of the concrete functional of form (1) we deal with is not taken into account we most often use the systems of functions

$$g_k(x) = x^{k-1}(x-a)(x-b) \qquad (k = 1, 2, 3, \ldots) \tag{4}$$

or

$$g_k(x) = \sin\frac{k\pi(x-a)}{(b-a)} \qquad (k = 1, 2, 3, \ldots) \tag{5}$$

It is readily shown that each of these systems is linearly independent and complete in C_1. It can also be proved that if $P(x) > 0$ and the original problem possesses a unique solution, the approximate solutions constructed according to the Ritz method converge to the exact solution as $n \to \infty$. (For instance, the uniqueness condition is sure to be fulfilled if $Q(x) \geqslant 0$.)

In concrete problems it is difficult to check the conditions of convergence theorems which thus only retain their theoretical significance. In real problems we more often use some theorems giving estimations for the discrepancy between the exact and approximate solutions. Some theorems of this kind concerning the Ritz method were for the first time proved in 1918 by N.M. Krylov (1879-1955), a prominent Soviet mathematician. But nevertheless these estimates sometimes are impractical in concrete situations and are only applicable to a narrow class of problems.

Therefore, in practical applications of the Ritz method we usually interpret the convergence of the approximations in pra-

ctical sense, that is try to choose a small number of coordinate functions such that it is possible to construct an approximate solution with a sufficient accuracy. To this end, we can try to use a number of first terms of system (4) or (5) or of some other systems of functions. If, when choosing an appropriate system of functions, we manage to take into account the concrete structure of functional (1), the convergence is usually accelerated. For instance, if $a=-b$ and the conditions of the problem indicate that the solution must be an even function, it is clear that when using system (4) we must limit ourselves to odd numbers k. On the other hand, if we do not know whether the solution is even or odd we cannot restrict the set of the coordinate functions to the even functions $g_k(x)$ because this is equivalent to imposing the condition that the solution is even, which may not be the case. If intuitive considerations indicate that the exact solution has a "peak" at a point $x=x_0$ it is preferable to choose one of the first functions $g_k(x)$ possessing the same peak and the like. Some further considerations concerning the choice of the system of coordinate functions will be given below.

When solving a variational problem we can look for the function giving the sought-for extremal or stationary value to the functional or try to compute this value itself. It is clear that the latter approach is more advantageous because the error in the determination of the stationary value is of higher order of smallness relative to the corresponding error in the extremizing function (cf. the problem of finding a stationary value of a function of one variable). Therefore it is preferable to reduce complicated problems (provided it is possible) to finding stationary values.

Let us return to determining an approximate solution of form (2). Substituting expression (2) into the right-hand side of (1), opening the brackets and performing integration we arrive at a polynomial of the second degree of the form

$$I\{g_0+\sum c_k g_k\}=V(c_1, c_2, \ldots, c_n) \tag{6}$$

in the unknown arbitrary parameters $c_1, c_2, \ldots, c_n$. It is obvious that if, for instance, the original problem involves the determination of a minimum, it is desirable to choose the values of the parameters c_k in such a way that expression (6) is minimized as well. The same applies to any stationary value of the functional: in the region of slow variation of functional (1) expression (6) must also vary slowly when the parameters are varied. Hence, we arrive at the system of equations

$$V'_{c_k}(c_1, c_2, \ldots, c_n)=0 \qquad (k=1, 2, \ldots, n) \tag{7}$$

which thus determines the values of the parameters c_k.

Let us introduce the notation

$$(f,\ g)_1 = \int_a^b [P(x) f'(x) g'(x) + Q(x) f(x) g(x)]\, dx$$

Then the quadratic form (in the variables c_k) entering into the right-hand side of expression (6) is written as

$$\sum_{k,\ l=1}^{n} (g_k,\ g_l)_1 c_k c_l \tag{8}$$

(check up this expression!). Consequently, (7) is a system of n linear algebraic equations in the unknowns $c_1, c_2, \ldots, c_n$ with symmetric matrix

$$2((g_k,\ g_l)_1) \qquad (k,\ l = 1,\ 2,\ \ldots,\ n)$$

which can easily be solved when n is not large. We note that if $P(x) > 0$, $Q(x) \geqslant 0$ and the functions $g_1, g_2, \ldots, g_n$ are linearly independent it can be readily shown (let the reader prove it!) that quadratic form (8) is positive definite, which facilitates the solution of system (7) (see the end of Sec. IV.4.3). To complete the application of the Ritz method we substitute the values of c_k thus found into the right-hand side of (2) and get the desired approximate solution of the original problem.

The matrix $((g_k,\ g_l)_1)$ in system (7) may have a large condition number (see Sec. IV.4.2), and then the rounding errors appearing in the process of solution and the errors in the initial data may yield a doubtfull final result; it can turn out that the increase of the number of coordinate functions in the Ritz method does not reduce the error of the approximate solution and even makes it increase. The matrix $((g_k,\ g_l)_1)$ is symmetric and has no negative eigenvalues; therefore the condition number of system (7) is equal to the ratio of the greatest eigenvalue of the coefficient matrix to the least one, and if the system degenerates (i. e. its matrix becomes singular) the condition number tends to infinity.

Here we have the best situation when

$$|(g_k,\ g_l)_1| = \delta_{kl} \tag{9}$$

because then the condition number is equal to unity. If the original system of coordinate functions does not possess this property it can be transformed by analogy with the well-known orthogonalization process (e.g. see [107], Sec. VII.21) in order to achieve condition (9). In particular, condition (9) is fulfilled for the functions

$$g_k(x) = |\lambda_k|^{-\frac{1}{2}} y_k(x) \qquad (k = 1,\ 2,\ \ldots)$$

where λ_k and $y_k(x)$ are, respectively, the eigenvalues and the eigenfunctions of the corresponding operator of type (2.26) which were constructed in Sec. 2.8 (prove this assertion!). If we cannot even approximately determine the eigenfunctions, it sometimes is expedient to use the eigenfunctions of a modified operator in which the coefficient $Q(x)$ is given an appropriate small variation because it can be shown that the eigenfunctions satisfy conditions (9) in the asymptotic sense. If this technique is impractical we can try to vary the leading coefficient $P(x)$ although this may result in a system with large condition number. By the way, in the particular case $P \equiv 1$, $Q \equiv 0$ the resultant system coincides, to within constant factors, with system (5) (why?).

These considerations concerning the choice of the system of coordinate functions are also applied to functionals involving functions of several independent variables.

In practical calculations, when it is necessary to estimate the accuracy of an approximate solution found by means of the Ritz method, we usually compare the numerical results obtained for a given system of coordinate functions and for an extended system (to which some other functions have been added) or for a new system of coordinate functions. If these results are close to each other we have good reason to expect that we have obtained a sufficiently accurate approximate solution because it is unlikely that this is a mere coincidence. (It should be noted that when the system of coordinate functions is extended we must compare the whole solutions but not their coefficients in the same coordinate functions because in the case of matrix (8) with large condition number a discrepancy in the values of these coefficients does not necessarily imply that the solutions themselves considerably differ.) If we have a series of single-type calculations this comparison is carried out for a few number of typical examples.

Consider the following example. Let it be required to find an extremal of the functional

$$I = \int_0^1 (y'^2 + y^2)\, dx \tag{10}$$

where the function $y(x)$ satisfies the boundary conditions

$$y(0) = 0, \quad y(1) = 1 \tag{11}$$

Here we readily write the exact solution

$$y_{\text{exact}} = \frac{\sinh x}{\sinh 1} \tag{12}$$

with which the approximate solutions can be compared. Let us put $g_0(x) = x$ in (2) and take as coordinate functions the first functions of system (4). To begin with, we take $n = 1$. This

means that we are looking for an approximate solution of the form

$$y = x + cx(1-x)$$

Substituting this expression into (10) and integrating we receive (check it up!)

$$I = \frac{11}{30}c^2 + \frac{1}{6}c + \frac{4}{3}$$

Equating to zero the derivative of I with respect to c we find the value $c = -\frac{5}{22}$ and thus obtain the approximate solution

$$y_{\text{I approximate}} = x - \frac{5}{22}x(1-x) + c_2x^2(1-x) \tag{13}$$

Now let us put $n = 2$ which leads to an approximate solution of the form

$$y = x + c_1x(1-x) + c_2x^2(1-x)$$

After it has been substituted into (10) we obtain (check up the result!)

$$I = \frac{11}{30}c_1^2 + \frac{11}{30}c_1c_2 + \frac{1}{7}c_2^2 + \frac{1}{6}c_1 + \frac{1}{10}c_2 + \frac{4}{3}$$

Equating to zero the partial derivatives of this expression with respect to c_1 and c_2 we arrive at the system of equations

$$\left.\begin{aligned} \frac{11}{15}c_1 + \frac{11}{30}c_2 + \frac{1}{6} = 0 \\ \frac{11}{30}c_1 + \frac{2}{7}c_2 + \frac{1}{10} = 0 \end{aligned}\right\}$$

whose solution is $c_1 = -\frac{69}{473} = -0.1459$, $c_2 = -\frac{7}{43} = -0.1628$, and therefore

$$\begin{aligned} y_{\text{II approximate}} &= x - 0.1459x(1-x) - 0.1628x^2(1-x) = \\ &= 0.8541x - 0.0169x^2 + 0.1628x^3 \end{aligned} \tag{14}$$

Computing the values of the exact solution and of the approximate solutions thus found we obtain the following results:

x	0.0	0.2	0.4	0.6	0.8	1.0
y_{exact}	0.0000	0.1713	0.3495	0.5417	0.7557	1.0000
$y_{\text{I approximate}}$	0.0000	0.1638	0.3457	0.5457	0.7638	1.0000
$y_{\text{II approximate}}$	0.0000	0.1714	0.3494	0.5415	0.7558	1.0000

The comparison shows that the simplest solution (13) which was found quite easily possesses the accuracy which is sufficient for most practical aims and that approximation (14) practically coincides

with the exact solution. Comparing the corresponding values of the functional I we obtain even more visual demonstration of the accuracy of the method:

$$I\{y_{\text{exact}}\} = \frac{\cosh 1}{\sinh 1} = 1.313, \quad I\{y_{\text{I approximate}}\} = \frac{347}{264} = 1.314$$

We see that the relative error of $I\{y_{\text{I approximate}}\}$ is 0.1%; to estimate the error of $I\{y_{\text{II approximate}}\}$ we must considerably increase the accuracy of the calculations. Note that for function (12) functional (10) attains its absolute minimum under boundary conditions (11) (why?), and therefore its values for the approximate solutions are naturally greater than the exact minimum value.

In the above example the exact solution is expressible in terms of elementary functions (but nevertheless approximate solution (13) is of interest since it is represented by simpler functions). But if, for instance, we take the functional

$$\int_0^1 (y'^2 + xy^2)\, dx$$

its extremals are no longer expressible as combinations of elementary functions. Since this functional is of the same type as (10) the expected accuracy of calculations is naturally the same as for functional (10).

If the boundary condition on one or both end points is of the form $y' = 0$ or of a more general form (2.37) we can simply discard this natural boundary condition when constructing approximate solutions but, on the other hand, to accelerate the convergence, it is advisable to use coordinate functions which satisfy homogeneous boundary conditions appearing for the exact solution. For example, if we have the boundary conditions $y(0) = 0$, $y'(l) = 0$ we can use the functions

$$g_1(x) = x - \frac{x^2}{2l}, \quad g_k(x) = x^{k-1}(l-x)^2 \qquad (k = 2, 3, 4, \ldots)$$

(let the reader explain why it is expedient to take the function $g_1(x)$ in this form).

By the way, in the case of arbitrary boundary conditions we sometimes follow the scheme in which these conditions are not taken into account when we choose coordinate functions (for instance, we can simply put $g_k(x) = x^{k-1}$, $k = 1, 2, \ldots$), and are replaced by some appropriate constraints imposed on the coefficients entering into the representation $y = \sum_{k=1}^{n} c_k g_k(x)$.

The application of the Ritz method to functionals involving derivatives of higher order, functionals of several functions and

to problems of determining conditional stationary values is performed in like manner and is also very effective.

2. Application to Boundary-Value Problems. The effectiveness of the Ritz approximate method gave rise to the following scheme for solving a boundary-value problem: given an ordinary differential equation and boundary conditions, we construct a functional whose Euler's equation is equivalent to the given equation and then apply the Ritz method. For instance, to choose a functional of form (1) for an equation of the form

$$a(x)\,y''+b(x)\,y'+c(x)\,y=f(x)$$

we multiply the equation by a factor $\mu(x)$ such that the equalities

$$-2P=a\mu,\quad -2P'=b\mu,\quad 2Q=c\mu,\quad -R=f\mu \tag{15}$$

hold. The first two equalities result in

$$P=\exp\int_{x_0}^{x}\frac{b}{a}\,dx,\quad \mu=-\frac{2}{a}P$$

after which we determine Q and R from the last two equalities (15).

For example, let the reader show that the **Bessel equation** (e.g. see [107], Sec. XV.26)

$$x^2y''+xy'+(x^2-p^2)\,y=0$$

is equivalent to Euler's equation corresponding to the functional

$$\int\left[xy'^2-\left(x-\frac{p^2}{x}\right)y^2\right]dx$$

If, for instance, this equation is considered on the interval $1\leqslant x\leqslant 2$, the application of the Ritz method yields a high accuracy if we only take one coordinate function of type (4) or (5).

An important feature of the method described above is that it can be used without writing explicitly the functional to be minimized. Indeed, remembering the method of deriving Euler's equation used in Sec. 1.7 and generalizing it we can suppose that the variation of the functional $I\{y\}$ which should be minimized by means of the Ritz method is representable in the form

$$\delta I\{y,\ \delta y\}=(L\{y\},\ \delta y) \tag{16}$$

where L is an operator. (For instance, in Sec. 1.7 we had $L\{y\}=F'_y(x,\ y,\ y')-\frac{d}{dx}F'_{y'}(x,\ y,\ y')$.) Then Euler's equation takes the form $L\{y\}=0$, and it is readily shown that

$$\frac{\partial}{\partial c_j}I\{g_0+\sum c_kg_k\}=(L\{g_0+\sum c_kg_k\},\ g_j)$$

Consequently, when we equate to zero these partial derivatives, which is necessary for the functional to be minimized, we obtain the same result as if we constructed an approximate solution of Euler's equation with the aid of formula (2) by means of the method of moments (e.g. see [107], Sec. XV.29) taken with respect to the functions $g_1, g_2, \ldots, g_n$. Therefore, it only remains to make one more step to seek the approximate solution of the given boundary-value problem for the equation $L\{y\} = 0$ in form (2) irrespective of any functional. The functions g_i should be chosen in accordance with the boundary conditions, and the coefficients c_i are found from the requirement that the moments must be equal to zero:

$$(L\{g_0 + \sum c_k g_k\},\ g_j) =$$
$$= \int_a^b L\left(g_0(x) + \sum_{k=1}^{n} c_k g_k(x)\right) g_j(x)\,dx = 0 \qquad (j = 1, 2, \ldots, n)$$

We thus arrive at the **Galerkin method** (also referred to as the **Bubnov-Galerkin method**). It was for the first time applied in 1913 to problems of the theory of elasticity by I. G. Bubnov (1872-1919), a Russian mechanician, and generalized in 1915 by B.G. Galerkin (1871-1945), a Soviet scientist. For Euler's equations encountered in the calculus of variations this method coincides with the Ritz method but it is also applicable to many ordinary and partial differential equations which are not Euler's equations and appear irrespective of variational problems. The Galerkin method is one of the most universal methods for solving boundary-value problems for these equations. The Soviet scientist G.I. Petrov (born in 1912) suggested in 1940 a modification of this method in which the solution of the problem is sought in the same form (2) but the moments which are equated to zero are taken with respect to another system of functions $h_1(x)$, $h_2(x), \ldots, h_n(x)$. For this method we refer the reader to [100] and [102].

3. Infinite Sequence of Coordinate Functions. It is sometimes possible to find the general formula for the coefficients c_k in representation (2) and then pass to the limit for $n \to \infty$ to obtain the sought-for solution in the form of an infinite series. In these circumstances we can at the very beginning take the infinite sequence of coordinate functions and construct the solution in the form of the sum of a series with undetermined coefficients.

For example, let it be necessary to minimize the integral

$$I\{y\} = \int_{-l}^{l} \left(\int_{-l}^{l} y'(s) \frac{ds}{x-s} \right) y(x)\,dx \qquad (17)$$

where the function $y(x)$ is subject to the integral constraint

$$J\{y\} = \int_{-l}^{l} y(x)\,dx = S = \text{const} \tag{18}$$

(the value of S is given) and to the boundary conditions

$$y(-l) = y(l) = 0 \tag{19}$$

The integral with respect to s in (17) is singular and is understood in the sense of Cauchy's principal value (e.g. see [107], Sec. XIV.19). This problem is posed in theoretical aerodynamics where the function $y(x)$ describes the dependence of the circulation of the air flux flowing round the wing of an airplane on the coordinate s reckoned along the wing, $2l$ is the wing span and the value of the functional $I\{y\}$ is proportional to the air drag (the compressibility of the air and friction are neglected). The integral $J\{y\}$ is proportional to the lift force, and thus we deal with the problem of determining the distribution of the circulation along the wing such that the aerodynamic drag has the minimum value for the given lift force.

To solve the problem we make the change of the independent variable $x = -l\cos\theta$; then the ends of the wing are determined by the values $\theta = 0$ and $\theta = \pi$. Taking into account boundary conditions (19), we shall seek the solution in the form

$$y = A_1 \sin\theta + A_2 \sin 2\theta + A_3 \sin 3\theta \tag{20}$$

Substituting (20) into (18) we obtain

$$\int_0^{\pi} (A_1 \sin\theta + A_2 \sin 2\theta + \ldots)\, l \sin\theta\, d\theta = S$$

whence

$$A_1 = \frac{2S}{\pi l}$$

(check up the result!). Now substituting expression (20) into (17) we receive

$$I\{y\} = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} A_j A_k k \int_0^{\pi} \left(\int_0^{\pi} \frac{\cos k\psi}{\cos\psi - \cos\theta}\, d\psi \right) \sin j\theta \sin\theta\, d\theta \tag{21}$$

The inner integral in (21) is equal to $\frac{\pi \sin k\theta}{\sin\theta}$ (see the next to the last formula in Sec. II.3.5). Substituting this value into (21) and taking into account the orthogonality of the sines we obtain

$$I\{y\} = \frac{\pi^2}{2} \sum_{k=1}^{\infty} k A_k^2$$

The coefficient A_1 has been already determined while the other coefficients A_k are arbitrary, and therefore it is clear that $I\{y\}$ attains its least value when $A_2 = A_3 = A_4 = \ldots = 0$, that is

$$y = \frac{2S}{\pi l} \sin\theta = \frac{2S}{\pi l^2} \sqrt{l^2 - x^2}$$

The last formula expresses the sought-for solution of the problem.

4. The Ritz Method for Functionals of Functions of Several Variables. The scheme of the Ritz method for functionals dependent on functions of several variables is essentially the same as for functions of one variable but the choice of coordinate functions is connected with some difficulties when the given boundary conditions must be fulfilled for a domain of a complex form. For the sake of simplicity, we shall consider functions of two independent variables. Suppose that the solution is constructed in a domain (G) with boundary (Γ), and the boundary condition has the form

$$u|_{(\Gamma)} = 0$$

Then, to form a system of coordinate functions, we try to find a continuous function $\omega(x, y)$ which is positive in the interior of (G) and equal to zero at all points of (Γ). If such a function has been chosen we can take as a system of coordinate functions the collection of the functions

$$g_{ij}(x, y) = x^i y^k \omega(x, y) \quad (j, k = 0, 1, 2, \ldots) \qquad (22)$$

numbered with two indices.

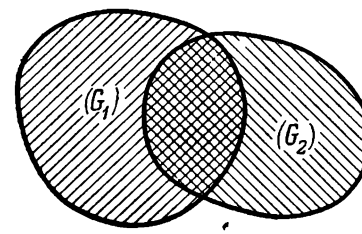


Fig. 109

In particular, if the domain (G) is determined by an inequality $F(x, y) > 0$ we can simply put $\omega \equiv F$. Combining functions of that type constructed for simple domains we can obtain coordinate functions for domains of a more complex shape. For example, if certain functions $\omega_1(x, y)$ and $\omega_2(x, y)$ are chosen, respectively, for two domains (G_1) and (G_2) (see Fig. 109), we can put

$$\omega = \omega_1(x, y)\,\omega_2(x, y)$$

for the *intersection* of (G_1) and (G_2) and

$$\omega = \begin{cases} \omega_1(x, y) & \text{if } (x, y) \in (G_1) \text{ and } (x, y) \notin (G_2), \\ \omega_2(x, y) & \text{if } (x, y) \in (G_2) \text{ and } (x, y) \notin (G_1), \\ \omega_1(x, y) + \omega_2(x, y) & \text{if } (x, y) \text{ belongs to the intersection of } (G_1) \text{ and } (G_2) \end{cases}$$

for their *union* (let the reader show that these functions are continuous). (The **intersection** or **product** or **meet** of two sets A and B is their common part denoted as $A \cap B$ or $A \cdot B$ or AB; the **union** of A and B is the set consisting of all the points of A and all the points of B and is denoted as $A \cup B$; also see Sec. IV.5.3.) We can similarly construct coordinate functions for the intersection and union of any number of domains.

If it is possible to take an appropriate coordinate system u, v in which the domain (G) is determined by inequalities $\alpha \leqslant u \leqslant \beta$, $\gamma \leqslant v \leqslant \delta$, we can take, as coordinate functions, expressions of the form

$$g_j(u)\, h_k(v) \qquad (j,\ k = 1,\ 2,\ 3,\ \ldots)$$

where $g_j(u)$ $(j = 1,\ 2,\ \ldots)$ is a system of coordinate functions for the interval $\alpha \leqslant u \leqslant \beta$ which vanish at its end points and $h_k(v)$, $k = 1,\ 2,\ \ldots$ is the same for the interval $\gamma \leqslant v \leqslant \delta$ (e.g. see an analogous construction in [107], Sec. XVII.30). In particular, for each interval we can take an appropriate system of eigenfunctions.

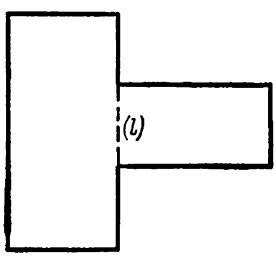


Fig. 110

For domains composed of two or more regions of simpler form (see Fig. 110) we can also apply the following technique: for the common part (l) of the boundaries of the constituent regions we choose an approximate formula for the solution which contains several parameters $c_1, c_2, \ldots, c_m$ and construct the solution by means of the scheme of the Ritz method for each constituent region separately imposing the condition that the functional should assume a stationary value with respect to all the parameters involved including $c_1, \ldots, c_m$.

Let us consider an example. Take the functional

$$I\{u\} = \int_{-a}^{a} dx \int_{-b}^{b} \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 + 2Au \right] dy \tag{23}$$

$(A = \text{const})$ dependent on the function $u(x, y)$ defined in the domain (G): $-a \leqslant x \leqslant a$, $-b \leqslant y \leqslant b$ with zero boundary values on the contour (Γ) bounding the domain. The problem of determining a stationary value of functional (23) is equivalent to the boundary-value problem

$$\nabla^2 u = A \quad \text{for } (x, y) \in (G),\ u|_{(\Gamma)} = 0. \tag{24}$$

Here we can put $\omega = (a^2 - x^2)(b^2 - y^2)$ and use system of functions (22) with even indices j and k because, as follows from the prin-

ciple of the maximum (see Sec. II.2.6), problem (24) possesses a unique solution and, besides, it is invariant with respect to the mappings $x \to -x$ and $y \to -y$, and therefore the solution must be an even function of x for any fixed y and an even function of y for any fixed x (the reader should carefully examine this argument). Let us take the simplest variant of the Ritz method with one coordinate function and look for an approximate solution of the form

$$u = c_1 (a^2 - x^2)(b^2 - y^2)$$

Substituting this expression into (23) we obtain (check it up!)

$$I = \frac{128}{45} a^3 b^3 (a^2 + b^2) c_1^2 + \frac{32}{9} a^3 b^3 A c_1$$

Now using the condition that the quantity I must assume a stationary value we find $c_1 = -\frac{5A}{8(a^2+b^2)}$, that is

$$u_{\text{approximate}} = -\frac{5A}{8(a^2+b^2)} (a^2 - x^2)(b^2 - y^2)$$

The comparison with a more accurate solution (which we do not give here) shows that for $a = b$ the maximum relative error of the solution found above estimated for different points of the domain (G) does not exceed 6% while its average value is 1%.

The next example deals with the functional

$$I\{u\} = \int_{(G)} \left[\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right] dG \qquad (25)$$

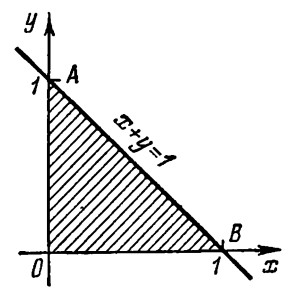


Fig. 111

Let it be required to construct the function $u(x, y)$ giving a stationary value to the functional in the domain (G) which is the right triangle depicted in Fig. 111 under the boundary condition

$$u|_{x+y=1} = x^2 + y^2$$

Integral (25) is called (for any number of independent variables and any domain (G)) the **Dirichlet integral** after P.G.L. Dirichlet (1805-1859), a prominent German mathematician. It is encountered in various divisions of mathematics; the Euler-Ostrogradsky equation corresponding to functional (25) is nothing but Laplace's equation $\nabla^2 u = 0$.

In this problem the boundary values are only set for the hypotenuse of the triangle (G), and therefore on the legs of the

triangle the solution must satisfy the natural boundary condition $\frac{\partial u}{\partial n}=0$.

To transform the problem to a homogeneous boundary condition we make the substitution

$$u=(x^2+y^2)+z$$

which turns functional (25) into

$$I\{z\}=\iint\limits_{(G)}\left[\left(\frac{\partial z}{\partial x}\right)^2+\left(\frac{\partial z}{\partial y}\right)^2+4x\frac{\partial z}{\partial x}+4y\frac{\partial z}{\partial y}\right]dG+\frac{2}{3}$$

while the boundary condition is replaced by

$$z|_{x+y=1}=0$$

As a system of coordinate functions we can take

$$g_{jk}(x, y)=[1-(x+y)]\,x^jy^k \quad (j, k=0, 1, 2, \ldots)$$

and construct an approximate solution of the form

$$z_{\text{approximate}}=[1-(x+y)]\sum_{j, k=0}^{n}a_{jk}x^jy^k \tag{26}$$

Expression (26) contains arbitrary parameters a_{jk}; it turns out that the properties of the solution of the problem make it possible to reduce the number of arbitrary parameters a_{jk} for any fixed n (and thus reduce the amount of calculations) or increase n for any fixed number of arbitrary parameters (that is to increase the accuracy of the approximation). Indeed, note that the problem is invariant with respect to the transformation $(x, y)\to$ $\to(y, x)$, and therefore we must have $a_{jk}=a_{kj}$. Besides, the exact solution satisfies the natural boundary condition $\frac{\partial u}{\partial n}=0$ on the legs of the triangle (G), which implies

$$\left.\frac{\partial z}{\partial x}\right|_{x=0}=0, \quad \left.\frac{\partial z}{\partial y}\right|_{y=0}=0 \tag{27}$$

and therefore it is natural to impose the same conditions on the approximate solution (26). This additional requirement accounts for what has been said about the arbitrary parameters and the accuracy of the approximation.

Now let us replace, for simplicity's sake, n by ∞. Then we have

$$\left.\frac{\partial}{\partial x}z_{\text{approximate}}\right|_{x=0}=\left.\left(-\sum_{j, k=0}^{n}a_{jk}x^jy^k+[1-(x+y)]\sum_{j, k=0}^{\infty}a_{jk}jx^{j-1}y^k\right)\right|_{x=0}=$$

$$=-\sum_{k=0}^{\infty}a_{0k}y^k+(1-y)\sum_{k=0}^{\infty}a_{1k}y^k=$$

$$= -\sum_{k=0}^{\infty} a_{0k} y^k + \sum_{k=0}^{\infty} a_{1k} y^k - \sum_{r=1}^{\infty} a_{1,r-1} y^r =$$

$$= (-a_{00} + a_{10}) + \sum_{k=1}^{\infty} (-a_{0k} + a_{1k} - a_{1,k-1})\, y^k$$

(where $k+1=r$) and, consequently, the first condition (27) yields

$$a_{10} = a_{00},\ a_{1k} = a_{0k} + a_{1,k-1} \quad (k = 1, 2, 3, \ldots) \tag{28}$$

Similarly, the second condition (27) results in

$$a_{01} = a_{00},\ a_{k1} = a_{k0} + a_{k-1,1} \quad (k = 1, 2, 3, \ldots) \tag{29}$$

Equalities (28) and (29) give us the required relations between the coefficients. By the way, if n is finite, we can only guarantee that conditions (27) are fulfilled approximately when equalities (28) and (29) hold, but nevertheless it is more advantageous to introduce these equalities than to discard them for $n < \infty$.

If, for instance, we put $n=1$, the approximate solution is written in the form

$$z_{\text{approximate}} = [1-(x+y)]\, a_{00}\, (1+x+y+2xy)$$

and thus involves only one arbitrary parameter instead of four; for $n=2$ we similarly obtain three arbitrary parameters instead of nine (check it up!).

In conclusion consider the functional

$$I\{u\} = \int_{-a}^{a} dx \int_{-b}^{b} [(\nabla^2 u)^2 - 2f(x, y)\, u]\, dy$$

whose Euler-Ostrogradsky equation is the nonhomogeneous biharmonic equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^2 u = f(x, y)$$

(see (3.70)) with the boundary conditions

$$u|_{(\Gamma)} = 0,\ \left.\frac{\partial u}{\partial n}\right|_{(\Gamma)} = 0$$

Here we can use the system of coordinate functions

$$(x^2-a^2)^2 (y^2-b^2)^2 x^j y^k \qquad (j, k = 0, 1, 2, \ldots) \tag{30}$$

or take the functions

$$g_{jk}(x, y) = \varphi_j(x)\, \psi_k(y) \tag{31}$$

where $\varphi_j(x)$, $j=1, 2, \ldots$, are the eigenfunctions of the corresponding one-dimensional problem

$$\frac{d^4 v}{dx^4} = \lambda v,\ \left. v\right|_{x=\pm a} = 0,\ \left.\frac{dv}{dx}\right|_{x=\pm a} = 0$$

and the functions $\psi_k(y)$, $k = 1, 2, \ldots$, are defined similarly. It is readily shown that we must necessarily have $\lambda > 0$ (to prove this inequality it is sufficient to multiply both sides of the equation by v, integrate them with respect to x from $-a$ to a and then twice perform integration by parts in the left-hand member). Consequently, the general solution of this equation expressing the eigenfunctions is written in the form

$$v = C_1 \cosh \sqrt[4]{\lambda}x + C_2 \sinh \sqrt[4]{\lambda}x + C_3 \cos \sqrt[4]{\lambda}x + C_4 \sin \sqrt[4]{\lambda}x$$

Using the boundary conditions we obtain (let the reader carry out the calculations) the following two types of eigenfunctions:

$$v = \frac{\cos \sqrt[4]{\lambda}a}{\cosh \sqrt[4]{\lambda}a} \cosh \sqrt[4]{\lambda}x - \cos \sqrt[4]{\lambda}x \quad \text{where} \quad \tanh \sqrt[4]{\lambda}a = -\tan \sqrt[4]{\lambda}a$$

and

$$v = \frac{\sin \sqrt[4]{\lambda}a}{\sinh \sqrt[4]{\lambda}a} \sinh \sqrt[4]{\lambda}x - \sin \sqrt[4]{\lambda}x \quad \text{where} \quad \tanh \sqrt[4]{\lambda}a = \tan \sqrt[4]{\lambda}a$$

(the eigenfunctions are defined to within arbitrary constant factors; in the above expressions these factors are chosen in such a

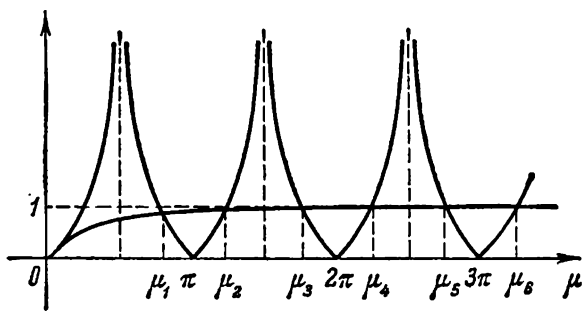


Fig. 112

way that the functions are of the order of unity for large values of λ). Putting $\sqrt[4]{\lambda}a = \mu$, i.e. $\lambda = \frac{\mu^4}{a^4}$, we see that the eigenvalues λ are found by means of the equation $\tanh \mu = \pm \tan \mu$ whose solution is demonstrated graphically in Fig. 112. We see that there is an infinite sequence of eigenfunctions; they can be easily computed with a desirable accuracy.

The calculations performed for the case $f \equiv \text{const}$, $a = b$ show that if we use system (30) with one arbitrary parameter the relative error of the approximate solution is less than 5% and if we take system (31) the error is less than 1.5%.

5. The Trefftz Method. If the Euler-Ostrogradsky equation is homogeneous, linear and has a sufficiently simple form (while the domain of integration can be bounded by a complex contour) the **Trefftz method** which we are going to discuss proves particularly effective. The idea of the method is that certain solutions of the Euler-Ostrogradsky equation are taken as coordinate functions and the coefficients entering into their linear combinations are determined from the boundary conditions. We shall demonstrate this method by the problem of finding a stationary value of the Dirichlet integral (25) for the functions u satisfying the boundary condition

$$u|_{(\Gamma)} = f, \tag{32}$$

where f is a given function defined on the contour (Γ). Let $u_0(x, y)$ be the unknown solution of the problem. We shall seek an approximate solution of the form

$$u_{\text{approximate}} = \sum_{k=1}^{n} c_k u_k(x, y) \tag{33}$$

where u_k are harmonic functions which are solutions of the Euler-Ostrogradsky equation $\nabla^2 u = 0$ (i.e. Laplace's equation) corresponding to functional (25). If the function u_0 were known we could pose the problem of determining the coefficients c_k in such a way that the function $u_{\text{approximate}}$ minimizes the integral

$$I\{u_{\text{approximate}} - u_0\} = \int_{(G)} \left[\left(\frac{\partial}{\partial x} u_{\text{approximate}} - \frac{\partial u_0}{\partial x} \right)^2 + \right.$$
$$\left. + \left(\frac{\partial}{\partial y} u_{\text{approximate}} - \frac{\partial u_0}{\partial y} \right)^2 \right] dG \tag{34}$$

If we replaced the function $u_{\text{approximate}}$ in (34) by an arbitrary function z (which is not necessarily of form (33)) and considered the problem of determining the function $z(x, y)$ minimizing the integral, the exact solution of the latter problem would be equal, to within an arbitrary constant summand, to the function u_0 (let the reader show that if z is an arbitrary function satisfying condition (32), integral (34) with $u_{\text{approximate}}$ replaced by z coincides with the difference $I\{z\} - I\{u_0\}$). Therefore, it appears natural to try to construct the approximate solution of form (33) minimizing functional (34). Taking the partial derivatives of (34) with respect to c_j, $j = 1, 2, \ldots, n$, and equating them to zero we arrive at the equations

$$\sum_{k=1}^{n} \int_{(G)} \left(\frac{\partial u_j}{\partial x} \frac{\partial u_k}{\partial x} + \frac{\partial u_j}{\partial y} \frac{\partial u_k}{\partial y} \right) dG\, c_k = \int_{(G)} \left(\frac{\partial u_j}{\partial x} \frac{\partial u_0}{\partial x} + \frac{\partial u_j}{\partial y} \frac{\partial u_0}{\partial y} \right) dG$$
$$(j = 1, 2, \ldots, n) \tag{35}$$

(check it up!). But it can be easily shown that for any harmonic function v and an arbitrary function u we have the formula

$$\int\limits_{(G)} \operatorname{grad} u \cdot \operatorname{grad} v \, dG = \int\limits_{(\Gamma)} u \frac{\partial v}{\partial n} d\Gamma$$

(To prove the formula it is sufficient to apply Ostrogradsky's theorem to the scalar field $u\nabla v$ and take advantage of formula (I.1.10).) Hence, on the basis of relation (32), we can rewrite equations (35) in the form

$$\sum_{k=1}^{n} \left(\int\limits_{(\Gamma)} \frac{\partial u_j}{\partial n} u_k d\Gamma \right) c_k = \int\limits_{(\Gamma)} \frac{\partial u_j}{\partial n} f \, d\Gamma \qquad (j = 1, 2, \ldots, n) \qquad (36)$$

which only involves the values of u_0 on the boundary (Γ) of the domain (G) which are equal to the values of the given function f. Thus, the resultant equations (36) can be used when the values of the function u_0 in the interior of the domain (G) are unknown. Resolving the system of equations (36) with respect to c_k and substituting the values of c_k thus found into (33) we obtain the same expression as if we minimized integral (34) and thus determine the sought-for approximate solution of the original problem.

Numerical calculations indicate that the Trefftz method is highly effective; it yields particularly accurate results when the harmonic functions entering into (33) are appropriately chosen in such a way that their behaviour imitates the properties of the sought-for solution u_0. It should be noted that this method is also applicable when boundary condition (32) is replaced by the condition $\left.\frac{\partial u}{\partial n}\right|_{(\Gamma)} = f$; in the latter case the right-hand side of (35) should be transformed into the integral $\int\limits_{(\Gamma)} u_j f d\Gamma$.

6. The Ritz Method for Computing Eigenvalues. The Ritz method can be used for solving eigenvalue problems. For example, the eigenvalue problem

$$-(P(x)y')' + Q(x)y = \lambda y, \quad y|_{x=a} = y|_{x=b} = 0$$

is equivalent to finding a function $y(x) \not\equiv 0$ satisfying the same boundary conditions and giving a stationary value to the functional

$$\int\limits_{a}^{b} [P(x)y'^2 + Q(x)y^2 - \lambda y^2] \, dx$$

Following the Ritz method we make substitution (2) with $g_0 \equiv 0$ and thus obtain the quadratic form

$$\sum_{j,k=1}^{n} a_{jk} c_j c_k - \lambda \sum_{j,k=1}^{n} b_{jk} c_j c_k$$

with the parameter λ and the given coefficients a_{jk}, b_{jk} connected by the relations $a_{jk}=a_{kj}$, $b_{jk}=b_{kj}$ $(j, k=1, 2, \ldots, n)$. The stationarity condition for the quadratic form is expressed by the equalities

$$\sum_k a_{jk}c_k-\lambda\sum_k b_{jk}c_k=0 \qquad (j=1, 2, \ldots, n) \tag{37}$$

Since we are interested in a nontrivial (i.e. nonzero) solution, the determinant of system (37) must be equal to zero:

$$\det((a_{jk})-\lambda(b_{jk}))=0 \tag{38}$$

The values of λ found from algebraic equation (38) serve as approximations to the eigenvalues of the original problem. Substituting these approximations into (37) we then determine c_k and thus obtain approximations for the corresponding eigenfunctions expressed by formulas (2).

The solutions of equations (38) give better approximations for the eigenvalues with smaller moduli while those with greater moduli are approximated less accurately. To improve the accuracy of the latter approximations we should increase n, which simultaneously makes it possible to estimate the accuracy achieved. But when the value of n is large it is more difficult to compute the values of determinant (38) (which is necessary for computing its zeros), and the amount of calculations particularly increases in the case of several independent variables. Therefore, it sometimes is more advantageous to compute these values with a comparatively large step for λ and then use the inverse interpolation (e.g. see [107], Sec. V.8).

Let us consider the following example:

$$-y''=\lambda y, \quad y|_{x=-1}=y|_{x=+1}=0 \tag{39}$$

We shall begin with showing that an eigenfunction of this problem is either even or odd. (The proof given below is based solely on the symmetry of the interval $-1\leqslant x\leqslant 1$ and on the fact that the coefficients $P(x)$ and $Q(x)$ are even functions and is irrelevant of the concrete structure of the equation.) We see that under the transformation $x\longrightarrow -x$ relations (39) go into themselves, while every eigenfunction is determined, for any given λ, to within a nonzero constant factor. Consequently, we must have $y(-x)\equiv ky(x)$. Substituting $-x$ for x into this identity we receive the relation $y(x)=ky(-x)=k^2y(x)$ whence $k=\pm 1$, which is what we set out to prove.

For this problem we can find the exact eigenvalues (see (2.35)) which are

$$\lambda_1=\frac{\pi^2}{4}=2.46740, \quad \lambda_2=\pi^2=9.8696, \quad \lambda_3=\frac{9\pi^2}{4}=22.21 \text{ etc.}$$

The functional corresponding to problem (39) has the form

$$I\{y\} = \int_{-1}^{1} (y'^2 - \lambda y^2)\, dx \tag{40}$$

We shall first take only the first function of system (4), that is put $y = a_0(1-x^2)$. Substituting this expression into (40) we obtain

$$I\{y\} = \frac{8}{3} a_0^2 - \frac{16}{15} \lambda a_0^2$$

In this case equation (38) is an equation of the first degree: $\frac{8}{3} - \frac{16}{15}\lambda = 0$; its solution is $\lambda_{1\ \text{approximate}} = 2.5$ which approximates λ_1 to an accuracy of 1.3%.

If we confine ourselves to determining even eigenfunctions the subsequent approximation should be taken in the form

$$y = (1-x^2)(a_0 + a_2 x^2)$$

Then (38) becomes an equation of the second degree. The computations (which we leave to the reader) yield

$$\lambda_{1\ \text{approximate}} = 2.46744, \quad \lambda_{3\ \text{approximate}} = 25.6$$

(Note that λ_2 corresponds to an odd eigenfunction.) We see that the relative errors of the approximations are 0.002% for λ_1 and 13% for λ_3. If we compute the next approximations for even eigenfunctions we obtain an error of $3 \cdot 10^{-7}\%$ for λ_1 and 0.5% for λ_3 while for λ_5 the error is of 30%. These data demonstrate the nonuniform character of the approximations mentioned above. The same tendency is found in the values of the eigenfunctions for which the errors are a little higher.

The approximate values obtained in the above problem are greater than the exact eigenvalues. It can be shown that this is a general property. The matter is that, for instance, λ_1 can be found as the solution of a minimum problem involving the class of all functions satisfying the boundary conditions and the corresponding normalization condition (see Sec. 2.8) while $\lambda_{1\ \text{approximate}}$ is the solution of the same problem for a narrower class of functions of a special form (namely, for the linear combinations of the coordinate functions), and, consequently, we always have $\lambda_{1\ \text{approximate}} \geqslant \lambda_1$. (Let the reader find out the conditions under which we have the equality sign in the latter relation.)

In conclusion we make one more remark which is of use for practical computations. Suppose that we are going to take the explicit expression of determinant (38) for a value of n which is neither large nor small (say, $n = 3$ or $n = 4$) in order to apply an iteration method for solving the resultant algebraic equation.

As is known, in this case it is preferable to have an accurate initial approximation. This approximation can be obtained for λ_1 if we equate to zero one of the principal minors (Sec. IV.2.3) of the matrix $(a_{jk})-\lambda(b_{jk})$, for instance, the first one. In fact, it is readily shown that this results in an approximate equation for λ_1 coinciding with the one found by the Ritz method in which the number of coordinate functions is appropriately reduced. An approximate expression for λ_2 can be obtained if we equate to zero the second principal minor and use the formula for the sum of the roots of an algebraic equation, etc.

7. The Ritz Method for General Functionals. The scheme of application of the Ritz method to functionals of a more general form (i.e. nonquadratic) is essentially the same as in Sec. 1 but the resultant systems of equations for the parameters are nonlinear, which leads to many complications. In this case approximate solutions are not necessarily looked for in the form of linear functions, that is we can use general formulas of the type

$$y=\varphi(x, c_1, c_2, \ldots, c_n) \tag{41}$$

(and analogous expressions for functions of several independent variables). The process of solving a nonlinear system of equations becomes extremely complicated when the number of the unknowns increases, and besides, it is preferable to begin with values which are as close as possible to the exact solution. Therefore, in problems involving general functionals a primary role belongs to intuition which sometimes makes it possible to construct a sufficiently accurate approximation to the sought-for solution by using a small number of parameters. If there is no information on the properties of the solution we can try to use approximate solutions of form (2), which at least guarantees a sufficient flexibility of the approximate expressions. As before, if the solutions corresponding to different values of n are close to one another we have good reason to expect that the approximations found are sufficiently accurate (an approximation corresponding to a smaller value of n can be taken as an initial approximation in the computations for a larger n, etc.), and the same refers to approximations corresponding to different systems of coordinate functions.

Suppose that we deal with a minimum variational problem. Then, after substitution (41) has been made, we arrive at a new problem of determining a minimum of a function dependent on a finite number n of independent variables $c_1, c_2, \ldots, c_n$. Such minimum problems for functions of n arguments appear not only in the calculus of variation but also in other divisions of mathematics, for example, in nonlinear programming (see Sec. IV.5.8). In modern computational mathematics there are a number of methods for solving these problems. Here we shall describe the

one known as the **gradient method** which proves effective in some simple cases. The independent variables will be denoted as x_1, x_2, ..., x_n.

Let us begin with the problem of finding an unconditional minimum for a given function $f(x_1, x_2, \ldots, x_n)$. As is known, the vector $-\text{grad } f(x)$ is at each point $x=(x_1, x_2, \ldots, x_n)$ in the direction of the fastest decrease ("steepest descent") of the function f. Consequently, the system of differential equations

$$\frac{dx_i}{dt} = -f'_{x_i}(x_1, x_2, \ldots, x_n) \qquad (i=1, 2, \ldots, n) \qquad (42)$$

specifies trajectories in the space x ($x=(x_1, \ldots, x_n)$) along which the function $f(x)$ decreases continuously, and, at least locally, the rate of decrease is the greatest. If the function remains bounded for all finite values of x a trajectory of this type either recedes to infinity or arrives at the vicinity of a point of local minimum. (Theoretically, a trajectory may happen to arrive at a point of minimax but, on the other hand, an infinitesimal variation of the initial conditions makes the trajectory change its behaviour and leave the vicinity of the minimax; practically, we can be sure that this case is avoided if the initial values are chosen at random.) Hence, if we find a trajectory which approaches a fixed point the latter is a point of minimum. If this local minimum coincides with the absolute one (see the end of Sec. IV.5.8) we thus receive the least value of the function f; if we have no information of that kind it is advisable to try to improve the result by constructing trajectories corresponding to various initial values taken at random.

The realization of the above scheme is connected with a complication due to the fact that when the moving point approaches a point of minimum its motion along the trajectory determined by system (42) is decelerated, and hence the point of minimum is only reached in the limit as $t \to \infty$. For instance, the equation $\frac{dx_1}{dt} = -2x_1$ is of form (42) for $n=1$, $f(x)=x_1^2$, and its solution is $x_1 = x_1^0 e^{-2(t-t_0)} \xrightarrow[t\to\infty]{} 0$. To avoid the deceleration we can introduce an appropriate factor, for instance, $\left(\sum_i (f'_{x_i})^2\right)^{-\frac{1}{2}}$, into the right-hand sides of (42), which does not change the trajectories but compensates for the deceleration (why?). We can also apply another technique: when the motion becomes sufficiently slow we proceed to solve the system of the equations $f'_{x_i}=0$ ($i=1, 2, \ldots, n$) by means of Newton's method.

Now suppose that it is required to determine a conditional minimum of a function $f(x_1, x_2, \ldots, x_n)$ under the k ($k<n$) in-

dependent finite constraints

$$\varphi_i(x_1, x_2, \ldots, x_n)=0 \qquad (i=1, 2, \ldots, k) \tag{43}$$

This means that the elementary displacements in the x-space $(x=(x_1, x_2, \ldots, x_n))$ specified by the differentials $dx_1, \ldots, dx_n$ must satisfy the relations

$$\sum_{j=1}^{n} \frac{\partial \varphi_i}{\partial x_j} dx_j=0 \qquad (i=1, 2, \ldots, k) \tag{44}$$

An elementary displacement for which the rate of decrease of the function f is the greatest is now obtained by projecting the vector $-\operatorname{grad} f$ onto the corresponding $(n-k)$-dimensional plane (hyperplane) determined by equations (44). A vector $\mathbf{b}$ going in this projected direction can be found, for instance, in the following way: we introduce the (column) number vectors

$$\mathbf{a}_i=\left(\frac{\partial \varphi_i}{\partial x_1}, \frac{\partial \varphi_i}{\partial x_2}, \ldots, \frac{\partial \varphi_i}{\partial x_n}\right)^* \qquad (i=1, 2, \ldots, k) \tag{45}$$

(the asterisk designates the operation of transposing the row vectors $\left(\frac{\partial \varphi_i}{\partial x_1}, \ldots, \frac{\partial \varphi_i}{\partial x_n}\right)$) which are perpendicular to the hyperplane (44), put

$$\mathbf{b}=-\operatorname{grad} f+\lambda_1 \mathbf{a}_1+\lambda_2 \mathbf{a}_2+\ldots+\lambda_k \mathbf{a}_k \tag{46}$$

and then choose the coefficients λ_i in such a way that the vector $\mathbf{b}$ becomes perpendicular to all the vectors $\mathbf{a}_i$ (let the reader carefully examine this construction). To determine λ_i, we take the scalar products of expression (46) by all vectors $\mathbf{a}_i$ and equate them to zero. This results in a system of equations of the first degree with unknowns λ_i; the determinant $D=\det(\mathbf{a}_i \cdot \mathbf{a}_j)$ of this system is called the **Gram determinant** (or, simply, the **Gramian**) for the vectors $\mathbf{a}_i$. It coincides with the determinant of the matrix of the quadratic form $(x_1\mathbf{a}_1+x_2\mathbf{a}_2+\ldots+x_k\mathbf{a}_k)^2$ and therefore is equal to the product of the eigenvalues of the matrix. Consequently, the Gramian is positive if the vectors $\mathbf{a}_i$ are linearly independent and is equal to zero if otherwise (why?). Thus, generally speaking, the condition $D \neq 0$ holds when relations (43) are independent (e.g. see [107], Sec. XI.15).

Hence, the directions of the displacements toward a point of conditional minimum of the function f for which the rate of motion is (locally) the greatest are determined by the system of differential equations

$$\frac{dx_i}{dt}=b_i(x_1, x_2, \ldots, x_n) \qquad (i=1, 2, \ldots, n) \tag{47}$$

where b_i $(i=1, 2, \ldots, n)$ is the projection of the vector $\mathbf{b}$ (formula (46)) on the x_i-axis. For small values of k the coefficients λ_i

can be found directly from the system of equations $\mathbf{b}\cdot\mathbf{a}_j=0$ $(j=1, 2, \ldots, k)$; for large values of k we can add, by analogy with Sec. IV.4.6, to (47) the system of k^2 equations for the elements of the matrix $(\mathbf{a}_i\cdot\mathbf{a}_j)^{-1}$. To insert corrections, we take some $n-k$ values from the values of x_i found for a chosen t, compute more accurate values of the remaining x_i's by applying Newton's method to equations (43) and then determine more accurately the elements of the matrix $(\mathbf{a}_i\cdot\mathbf{a}_j)^{-1}$ with the aid of the method of Sec. IV.4.6. The deceleration of convergence in the vicinity of the point of minimum can be avoided as in the case of an unconditional minimum. It should be noted that when applying Newton's method to a conditional extremum problem we should perform linearization not only with respect to the unknowns x_i but also with respect to Lagrange's multipliers λ_i (e.g. see [107], Sec. XII.10). The initial values of these quantities which are then corrected can be found by solving the differential equations because formulas (45) and (46) indicate that at the point of minimum, for $\mathbf{b}=0$, the values of λ_i thus found are nothing but Lagrange's multipliers.

If it is known that among the values of x_i we deal with there are k variables x_i, for instance, x_{n-k+1}, x_{n-k+2}, $\ldots$, x_n, which can be expressed as differentiable functions of the remaining x_i's, which, for example, is the case when

$$\frac{D(\varphi_1, \varphi_2, \ldots, \varphi_k)}{D(x_{n-k+1}, x_{n-k+2}, \ldots, x_n)}\neq 0$$

we can construct the trajectory of steepest descent in the space $x_1, x_2, \ldots, x_{n-k}$ without introducing Lagrange's multipliers. If this Jacobian becomes close to zero we can try to choose another collection of $n-k$ variables x_i, etc.

If the original problem requires to find a stationary value of the function f (instead of its minimum) we can apply the gradient method to the function $\varphi=\sum_i f'^2_{x_i}$. But it should be taken into account that the function φ may have points of minima different from the stationary points of the function f (let the reader investigate, in this connection, the case $n=1$), and therefore it may sometimes be useful to apply the gradient method repeatedly by taking a number of randomly chosen initial points.

In conclusion we note that, at least in principle, the gradient method can be directly applied to the given functional without using its finite-dimensional approximation. Indeed, suppose that we managed to represent the variation of a functional $I\{y\}$ in form (16). Then the operator L plays the role of the gradient of the functional $I\{y\}$, and the equation specifying the direction of

steepest descent takes the form

$$\frac{\partial y}{\partial t} = -L\{y\}$$

Here we have $y = y(x, t)$, and consequently this is a partial differential equation which must be solved for a chosen initial condition $y|_{t=0} = y_0(x)$ and the boundary conditions imposed on the functional I. This can also be performed practically but involves some sophisticated techniques and again involves finite-dimensional approximations. In the present course we shall not deal with partial differential equations.

8. Method of Least Squares. As was mentioned in Sec. 1, the effectiveness of the Ritz method makes it possible to develop an approximate method for solving differential equations based on reducing the original boundary-value problem to a variational one which then is solved with the aid of the Ritz method. The simplest technique of this kind (although not always the most effective) is based on the minimization of the mean square deviation (sometimes taken with a weight function). This is the so-called **method of least squares** (e.g. see [107], Sec. XV.29) which can be applied not only to differential equations but also to other types of equations involving one or several unknown functions.

For instance, take the problem of integrating the equation

$$y' = f(x, y) \quad a \leqslant x \leqslant b \tag{48}$$

with the initial condition

$$y(a) = y_a \tag{49}$$

where y_a is given. This problem is obviously equivalent to the minimization of the functional

$$I\{y\} = \int_a^b [y' - f(x, y)]^2 \, dx \tag{50}$$

in which the function $y(x)$ satisfies condition (49) at the end point $x = a$ while the end point value at $x = b$ is not subject to any constraint. The latter problem can be solved with the aid of the Ritz method.

A feature of this approach is that Euler's equation corresponding to functional (50) is written as

$$2[y' - f(x, y)][-f'_y(x, y)] - \frac{d}{dx} 2[y' - f(x, y)] = 0 \tag{51}$$

and reduces to the differential equation

$$y'' = f'_x(x, y) + f(x, y) f'_y(x, y) \tag{52}$$

of the second order which can be obtained by the differentiation of the original equation (48). The general solution of (52) involves two arbitrary parameters while the original equation (48) possesses a one-parameter general solution. Therefore, to determine the values of the arbitrary constants in the general solution of (52) we must resort to the natural boundary condition for the end point $x=b$ which has the form $[y'-f(x, y)]_{x=b}=0$. The latter condition enables us to pass from (52) to (48) (let the reader carry out the calculations by taking equation (52) in the equivalent form (51)).

The increase of the order of the derivatives demonstrated in the above problem is a demerit of the method of least squares although it is applicable to a wide class of problems. Therefore, if a given differential equation of the $2k$th order can be rewritten as Euler's equation corresponding to a functional J involving derivatives of order not higher than k it is advisable to apply the Ritz method to the functional J without using the method of least squares which involves a functional with derivatives of the $2k$th order. The matter is that the higher the order of the derivatives entering into a functional, the greater the error of the approximations obtained by means of the Ritz method.

In the above problem we can also minimize the functional

$$\int_a^b [y'-f(x, y)]^2\,dx+[y(a)-y_a]^2$$

(instead of (50)) considered without condition (49). This simple remark suggests the following modification of the method of least squares applicable to functions of any number of independent variables, any domain of integration, equation and boundary conditions. Let it be necessary to solve an equation of the form $M[u]=0$ for a function u defined in a domain (G) with boundary (Γ) and satisfying the boundary conditions

$$\mu_j[u]|_{(\Gamma)}=0 \qquad (j=1, 2, \ldots, k)$$

where $M[u]$ and $\mu_j[u]$ $(j=1, 2, \ldots, k)$ are some (differential) operators. (It is obvious that there is a wide variety of problems for which the corresponding equations and boundary conditions can be written in the above operator form after all the terms are transposed to the left-hand sides.) Then the sought-for solution minimizes the functional

$$a_0 \int_{(G)} (M[u])^2\,dG+\sum_{j=1}^{k} a_j \int_{(\Gamma)} (\mu_j[u])^2\,d\Gamma \tag{53}$$

where $a_0, a_1, \ldots, a_k$ is an arbitrary collection of positive constants whose introduction may prove useful when it is necessary

to balance, in a certain sense, the contributions made by the equation and the boundary conditions. (By the way, for the sake of simplicity we can simply put $a_j = 1$, $j = 0, 1, 2, \ldots, k$.) Functional (53) admits of the application of the Ritz method with an arbitrary system of coordinate functions (for instance, we can take the system of rational minomials with integral powers of the variables) because it is considered without boundary conditions. In particular, if the equation $M[u] = 0$ is simple (for instance, coincides with Laplace's equation), and it is possible to determine a family of its solutions involving parameters, we can use this family for the application of the Ritz method provided the combinations of the functions of the family are sufficiently flexible in the sense that they yield accurate approximations to the sought-for solution. Then the problem is essentially simplified because in this case functional (53) will only retain the integrals over the contour (Γ).

The above scheme is in an obvious manner generalized to systems of equations containing several unknown functions.

9. Method of L.V. Kantorovich. Here we shall give an example of the application of the method inaugurated in 1933 by L.V. Kantorovich. This is a generalization of the Ritz method to functionals of functions dependent on several arguments. Its basic idea is that instead of constant coefficients in coordinate functions some unknown functions of one independent variable are taken. After a linear combination of coordinate functions with these coefficients is substituted into the functional the condition of the stationarity of the resultant expression is reduced to a boundary-value problem for a system of differential equations because this expression is a functional, and we thus deal with an auxiliary variational problem. Solving this boundary-value problem we finally arrive at an approximate solution of the original problem. A boundary-value problem can rarely be solved analytically but in many cases it admits of the application of modern computational means. For instance, this is the case when we deal with a linear boundary-value problem, which corresponds to a quadratic functional. A linear combination of coordinate functions with functional coefficients is more flexible in the sense that it usually yields a more accurate approximation to the exact solution compared with the ordinary scheme of the Ritz method.

We shall demonstrate the method of L.V. Kantorovich by the example of the functional

$$I\{u\} = \int\limits_{(G)} \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 - 2u\right] dG \tag{54}$$

Let it be required to determine a stationary value of the

functional under the boundary condition

$$u|_{(\Gamma)}=0 \tag{55}$$

where (G) is the domain depicted in Fig. 113. We shall take the coordinate function $g_1(x, y)=y(x-y)$ which is positive everywhere in the domain (G) except the straight lines $y=x$ and $y=0$ where it vanishes. Following the method of Kantorovich we shall seek an approximate solution of the problem in the form

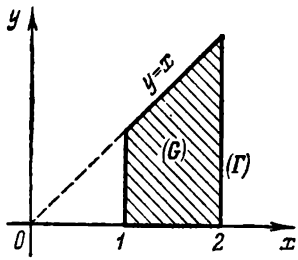


Fig. 113

$$u=a(x)\,g_1(x, y)=a(x)\,y(x-y) \tag{56}$$

where $a(x)$ is an arbitrary function which is only subject to the boundary conditions

$$a(1)=0, \quad a(2)=0 \tag{57}$$

Then the product (56) satisfies boundary condition (55). If we used the ordinary scheme of the Ritz method it would be natural to take the function $a(x)=c(x-1)(x-2)$ in which only the proportionality factor c is varied; in contrast to it, in the method of Kantorovich $a(x)$ is an arbitrary function satisfying relations (57) which is found from the condition that it gives a stationary value to functional (54) considered for the class of functions of form (56).

Substituting (56) into (54) and carrying out the calculations we obtain (check up the result!)

$$I\{u\}=\int_1^2\left(\frac{1}{30}x^5a'^2+\frac{1}{6}x^4aa'+\frac{2}{3}x^3a^2-\frac{1}{3}x^3a\right)dx$$

Euler's equation corresponding to the latter functional is written as

$$x^2a''+5xa'-10a=-5$$

and its general solution is readily expressed in terms of elementary functions (e.g. see [107], Sec. XV.19):

$$a=\frac{1}{2}+c_1x^{\sqrt{14}-2}+c_2x^{-\sqrt{14}-2}$$

Using conditions (57) we find the corresponding values of the arbitrary constants c_1 and c_2, which leads to the solution

$$a=\frac{1}{2}\left(1-\frac{4-2^{-\sqrt{14}}}{2^{\sqrt{14}}-2^{-\sqrt{14}}}x^{\sqrt{14}-2}-\frac{2^{\sqrt{14}}-4}{2^{\sqrt{14}}-2^{-\sqrt{14}}}x^{-\sqrt{14}-2}\right)=$$

$$=\frac{1}{2}(1-0.2951x^{1.742}-0.7049x^{-5.742})$$

Substituting this expression of $a(x)$ into (56) we obtain the sought-for approximate solution of the original problem.

If it is necessary to find a more accurate approximation we can take, instead of (56), the sum

$$u = a(x)\, y(x-y) + b(x)\, y^2 (x-y) \tag{58}$$

After the latter expression has been substituted into (54) we receive a system of two linear differential equations with two unknown functions $a(x)$ and $b(x)$. Besides, we can use the so-called **collocation method** (e.g. see [107], Sec. XV.29) whose application to problems of that kind was also suggested by L.V. Kantorovich. In the case under consideration we apply this method to the part $y=0$ of the contour (Γ), which is accomplished as follows. The Euler-Ostrogradsky equation corresponding to functional (54) has the form

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -1$$

Since the exact solution must satisfy this equation it is natural to impose the requirement that approximate solution (58) should satisfy this equation at least for $y=0$. This yields the relation

$$[a''(x)\, y(x-y) + 2a'(x)\, y - 2a(x) + b''(x)\, y^2 (x-y) + \\ + 2b'(x)\, y^2 + b(x)\,(2x-3y) + 1]_{y=0} = 0$$

whence $b = \frac{1}{2x}(2a-1)$. Substituting the last expression into (58) we thus see that the approximate solution should be looked for in the form

$$u = -\frac{y^2}{2x}(x-y) + \frac{y}{x}(x^2 - y^2)\, a(x) \tag{59}$$

Finally, substituting (59) into (54) and applying the stationarity condition we derive the differential equation for $a(x)$, etc. It appears natural to expect that the approximation of form (59) gives a better accuracy than the one of form (56) because the former possesses the same properties of the exact solution which are common to the latter and, besides, satisfies the above differential equation with respect to the variable x for the fixed value $y=0$ of the other independent variable.

If the domain of integration is a rectangle it is expedient to construct an approximate solution of the form $u = a(x)\, b(y)$ involving two unknown functions $a(x)$ and $b(y)$. After Euler's equations for $a(x)$ and $b(y)$ are found (let the reader write them down), we can apply the iteration method to them.

The reader can find many examples of the application of the method of Kantorovich in [58].

10. Euler's Method. When considering theoretical questions in Sec. 1.12 we used a discrete scheme in which functional (1.18) was approximately replaced by sum (1.61) involving a finite number of discrete values $\bar{y}_k$ of the unknown function. Euler's method we are going to discuss here consists in using this scheme as a numerical method.

Taking the stationarity condition for sum (1.61) we determine the values y_k which can naturally be regarded as approximate values of the sought-for solution at the nodes. If it is required to find the values of the function at other points we can use these values and apply the interpolation method. The accuracy of the method can be still improved if we apply more sophisticated formulas of numerical differentiation and integration than those used for deriving expression (1.61).

In [131] there is an instructive example of the application of the method which we shall give here. Let us consider the minimum problem for the functional

$$\int_0^1 (y'^2 + y^2 + 2xy)\, dx$$

in which the unknown function $y(x)$ satisfies the boundary conditions $y(0) = y(1) = 0$. Making use of formula (1.61) with $n = 5$ we obtain the expression

$$I_5 = 0.2\left\{\left[\left(\frac{y_1 - 0}{0.2}\right)^2 + 0^2 + 2\cdot 0\cdot 0\right] + \left[\left(\frac{y_2 - y_1}{0.2}\right)^2 + y_1^2 + 2\cdot 0.2 y_1\right] + \right.$$
$$+ \left[\left(\frac{y_3 - y_2}{0.2}\right)^2 + y_2^2 + 2\cdot 0.4 y_2\right] + \left[\left(\frac{y_4 - y_3}{0.2}\right)^2 + y_3^2 + 2\cdot 0.6 y_3\right] +$$
$$\left. + \left[\left(\frac{0 - y_4}{0.2}\right)^2 + y_4^2 + 2\cdot 0.8 y_4\right]\right\}$$

The stationarity conditions result in the relations

$$\left.\begin{aligned}
\frac{\partial I_5}{\partial y_1} &= 0.2\left\{\frac{2y_1}{0.2^2} - \frac{2(y_2 - y_1)}{0.2^2} + 2y_1 + 2\cdot 0.2\right\} = 0 \\
\frac{\partial I_5}{\partial y_2} &= 0.2\left\{\frac{2(y_2 - y_1)}{0.2^2} - \frac{2(y_3 - y_2)}{0.2^2} + 2y_2 + 2\cdot 0.4\right\} = 0 \\
\frac{\partial I_5}{\partial y_3} &= 0.2\left\{\frac{2(y_3 - y_2)}{0.2^2} - \frac{2(y_4 - y_3)}{0.2^2} + 2y_3 + 2\cdot 0.6\right\} = 0 \\
\frac{\partial I_5}{\partial y_4} &= 0.2\left\{\frac{2(y_4 - y_3)}{0.2^2} + \frac{2y_4}{0.2^2} + 2y_4 + 2\cdot 0.8\right\} = 0
\end{aligned}\right\} \qquad (60)$$

This is a system of linear algebraic equations which is readily solved. We eliminate y_1 from the first two equations, and then from the resultant equation and the third equation we eliminate y_2. Finally, from the equation obtained after the last elimination and from the fourth equation we eliminate y_3 and determine y_4 from

the remaining equation in one unknown. Then we substitute the value of y_4 into the fourth equation of system (60) and determine y_3; knowing y_3 and y_4 we find y_2 from the third equation of the system and, in conclusion, determine y_1 from the second equation. An elimination scheme of that type can always be applied to a diagonal system of linear equations and to a wide class of nondiagonal systems of arbitrary order with a coefficient matrix (a_{ij}) in which only diagonal elements and elements lying near the principal diagonal (for $|i-j| \leqslant 1$ or $|i-j| \leqslant 2$ and the like) are different from zero (in the above example we just had $|i-j| \leqslant 1$).

In this example we can easily construct the exact solution of Euler's equation and therefore we can present the following table of the exact values of the solution and the approximate values obtained after system (60) has been solved:

x	0.0	0.2	0.4	0.6	0.8	1.0
$-10^4 \cdot y_{\text{approximate}}$	0	286	503	580	442	0
$-10^4 \cdot y_{\text{exact}}$	0	287	505	583	444	0

Thus, we have received sufficiently accurate results. On the other hand, from the computational point of view this example only provides a pattern that can be followed in more complicated problems because it is easily seen that the same computational procedure is obtained if we simply pass from the variational problem to Euler's equation and then apply the ordinary finite difference method to solving the boundary-value problem for this equation.

CHAPTER VII

Integral Equations

The theory of **integral equations** deals with equations in which an unknown function occurs under an integral sign. It constitutes an important division of mathematical analysis and has many theoretical and practical applications. Although some integral equations were considered as early as the first half of the 19th century, their systematic theory was developed at the end of the 19th and the beginning of the 20th century by V. Volterra (1860-1940), an Italian mathematician, E.I. Fredholm (1866-1927), a Swedish mathematician, D. Hilbert (1862-1943), a famous German mathematician, and other scientists.

On the theory of integral equations we refer the reader to general courses [86], [98], [99], [110], [121], [130], [131] and [139].

§ 1. INTRODUCTION

1. Examples. As is known (e.g. see [107], Sec. XIV.26), if an external action applied to a linear system is described by a function $f(x)$ $(a \leqslant x \leqslant b)$, the result of the action (the response, "output" of the system) is described by the function

$$\hat{f}(x) = \int_a^b G(x, \xi) f(\xi)\, d\xi \tag{1}$$

where $G(x, \xi)$ is the influence function (Green's function) which is specified by the given system. For instance, $f(x)$ and $\hat{f}(x)$ may describe, respectively, the density of a load distributed along a bar and the corresponding deflection of the bar.

Suppose that the external action is unknown, whereas the response of the system corresponding to this action is given. Let it be necessary to reconstruct the external action from this output. Then the function $\hat{f}(x)$ in relation (1) as well as $G(x, \xi)$ are given, and $f(x)$ is the function to be found. Thus we arrive at an integral equation in $f(x)$.

There are cases when we know a linear combination $\varphi(x)=$ $=\alpha f(x)+\beta\hat{f}(x)$ of functions describing an external action and the corresponding output. Then, to reconstruct the external action we must solve the integral equation

$$\alpha f(x)+\beta\int_a^b G(x,\ \xi)\,f(\xi)\,d\xi=\varphi(x) \tag{2}$$

Let us take another problem reducible to an equation of type (2). Consider the equation of forced transversal vibrations of a string fixed at the end points $x=0$ and $x=l$ (see (VI.3.59)):

$$\rho u_{tt}''=Pu_{xx}''+f(x,\ t) \tag{3}$$

If we have a harmonic external action

$$f(x,\ t)=\varphi(x)\cos\omega t$$

and we are interested in a forced oscillation with the same frequency ω, i.e.

$$u(x,\ t)=v(x)\cos\omega t$$

then, after substituting this expression into (3) we arrive at the boundary-value problem

$$Pv''=-\rho\omega^2 v-\varphi(x)\qquad(0\leqslant x\leqslant l),\ v(0)=0,\ v(l)=0$$

If $\omega=0$, we deal with the problem of finding a stationary deflection of the string under an external action. The solution of this problem is given by the formula (e.g. see [107], Sec. XV.16)

$$v(x)=\int_0^l G(x,\ \xi)\left[-\frac{1}{P}\,\varphi(\xi)\right]d\xi \tag{4}$$

where the influence function of the problem is determined by the relation

$$G(x,\ \xi)=\begin{cases}-(l-\xi)\,xl^{-1} & \text{for}\quad 0\leqslant x\leqslant\xi\leqslant l\\ -\xi\,(l-x)\,l^{-1} & \text{for}\quad 0\leqslant\xi\leqslant x\leqslant l\end{cases}$$

But if $\omega\neq 0$ then, besides the load $\varphi(x)$, we have the inertia term $\rho\omega^2 v(x)$ dependent on the sought-for solution. Taking advantage of solution (4) we derive the relation

$$v(x)=-\frac{\rho\omega^2}{P}\int_0^l G(x,\ \xi)\,v(\xi)\,d\xi-\frac{1}{P}\int_0^l G(x,\ \xi)\,\varphi(\xi)\,d\xi$$

The function $v(x)$ being unknown, we have thus obtained an integral equation belonging to the same type as (2).

Note that a differential equation is a relationship between the values of given and unknown functions on an infinitesimal interval whereas an integral equation involves a finite interval. Therefore, it appears natural that the investigation of boundary-value problems is connected with integral equations.

2. Basic Classes of Integral Equations. An integral equation of the form

$$\int_a^b K(x, \xi)\, u(\xi)\, d\xi = f(x) \qquad (a \leqslant x \leqslant b) \tag{5}$$

with an unknown function $u(x)$ is called **Fredholm's integral equation of the first kind** and the function $K(x, \xi)$ is the **kernel** (or **nucleus**) of the equation. An equation of the form

$$u(x) = \int_a^b K(x, \xi)\, u(\xi)\, d\xi + f(x) \tag{6}$$

is known as **Fredholm's integral equation of the second kind.** It is obvious that in the general case (5) and (6) are nonhomogeneous linear equations. For $f \equiv 0$ we deal with the corresponding homogeneous equations.

If the integral in equations (5) and (6) is taken from a to x we refer to them, respectively, as **Volterra's equation of the first kind** and **Volterra's equation of the second kind.** By the way, if we formally put $K(x, \xi) \equiv 0$ for $\xi > x$ in Volterra's equation it can be regarded as a special case of Fredholm's equation. **Equations with real symmetric kernels** (for which $K(x, \xi) = K(\xi, x)$) form another important subclass of Fredholm's equations. For the case of a complex kernel the symmetry condition is taken in the form $K(x, \xi) = [K(\xi, x)]^*$. Such kernels are called **Hermitian** (this terminology is analogous to that of the theory of matrices; see Sec. IV.1.3). We also consider equations with kernels of the form $K(x, \xi) = T(x - \xi)$ (called **difference kernels**) which depend solely on the difference of the arguments.

We shall begin with the simplest case when a and b are finite and the kernel $K(x, \xi)$ is a continuous function. But the basic properties of solutions which are established below can also be easily proved for square-summable kernels satisfying the inequality

$$\int_a^b \int_a^b |K(x, \xi)|^2\, dx\, d\xi < \infty \tag{7}$$

Therefore, many authors speak of the Fredholm and Volterra equations whenever condition (7) holds.

For unbounded kernels an important case is when the kernel turns into infinity for $x = \xi$. Therefore, we specify the class of

equations with kernels having **polar singularity** which, for finite a and b, satisfy the condition

$$|K(x, \xi)| \leqslant \frac{\text{const}}{|x-\xi|^{\alpha}} \qquad (0<\alpha<1) \tag{8}$$

(it is meant, of course, that α is the least positive number for which inequality (8) holds as $x-\xi \to 0$). If $0<\alpha<\frac{1}{2}$, condition (7) is fulfilled (why?), i.e. the equation is of the Fredholm type. If $\alpha=1$ integral equations (5) and (6) with such a kernel are termed **singular**, and the integral is understood in the sense of Cauchy's principal value (e.g. see [107], Sec. XIV.19).

The theory of nonlinear equations is by far not so complete as the theory of linear equations. Some types of nonlinear equations will be considered in § 6.

The above definitions are directly extended to the systems of m integral equations with m unknown functions and also to equations with unknown functions of n independent variables. In the latter case an analogue of equation (6) can be written in the form

$$u(M) = \int_{(G)} K(M, N)\, u(N)\, d_N G + f(M) \tag{9}$$

where M and N are points of an n-dimensional domain (or an n-dimensional manifold) (G), and the subscript N in the differential indicates that the integration is performed with respect to the coordinates of the point N. Equation (9) can also be written in the coordinate form with an n-fold multiple integral. In the case of n independent variables condition (8) is replaced by the inequality

$$|K(M, N)| \leqslant \frac{\text{const}}{[\rho(M, N)]^{\alpha}} \qquad (0<\alpha<n)$$

where $\rho(M, N)$ is the distance between the points M and N.

We also consider integral equations which involve, instead of ordinary integrals (that is Riemann's and Lebesgue's integrals), integrals with respect to a measure (Stieltjes' integrals; e.g. see [107], Sec. XVI.19):

$$u(M) = \sum_j \int_{(G)} K_j(M, N)\, u(N)\, d_N \mu_j + f(M) \tag{10}$$

For instance, if equation (10) involves Lebesgue's measure and a measure concentrated at a point $M_0 \in (G)$ it takes the form

$$u(M) = \int_{(G)} K(M, N)\, u(N)\, d_N G + L(M)\, u(M_0) + f(M)$$

3. More on the Hilbert Space. In Sec. VI.1.3 we have mentioned the Hilbert space $L_2[a, b]$ (where a and b are not necessarily finite) of square-summable functions with norm (VI.1.11). For definiteness, we shall take the real space, that is we shall consider functions assuming real values. The complex Hilbert space $L_2[a, b]$ is introduced quite analogously, the norm being defined by the formula

$$\|\mathbf{f}\| = \sqrt{\int_a^b |f(x)|^2 dx}$$

Here $\mathbf{f}$ designates a function $f(x)$ regarded as an element (a generalized vector) of a linear space.

As is known (e.g. see [107], Sec. XVII.28), the Hilbert space $L_2[a, b]$ is an infinite-dimensional Euclidean space with the scalar product

$$(\mathbf{f}, \mathbf{g}) = \int_a^b f(x) g(x) dx \qquad (\mathbf{f}, \mathbf{g} \in L_2[a, b])$$

Among the properties of the space $L_2[a, b]$, the inequality

$$\left(\int_a^b f(x) g(x) dx\right)^2 \leqslant \int_a^b f^2(x) dx \int_a^b g^2(x) dx \tag{11}$$

is of great importance. It can be rewritten in the equivalent form

$$|(\mathbf{f}, \mathbf{g})| \leqslant \|\mathbf{f}\| \|\mathbf{g}\|$$

Let us prove that if a kernel $K(x, y)$ satisfies condition (7) and $u(x) \in L_2[a, b]$ then

$$\int_a^b K(x, \xi) u(\xi) d\xi \in L_2[a, b] \tag{12}$$

Indeed, applying inequality (11) we obtain

$$\int_a^b \left(\int_a^b K(x, \xi) u(\xi) d\xi\right)^2 dx \leqslant \int_a^b \left(\int_a^b [K(x, \xi)]^2 d\xi \int_a^b u^2(\xi) d\xi\right) dx =$$

$$= \int_a^b \int_a^b [K(x, \xi)]^2 dx d\xi \int_a^b u^2(\xi) d\xi \tag{13}$$

which is what we set out to prove.

Thus, if we denote the left-hand member of relation (12) by $Au(x)$, the symbol A can be interpreted as a linear operator which maps the space $L_2[a, b]$ into itself. Extracting the square

roots of both members of (13) we obtain the inequality

$$\|\mathbf{Au}\| \leqslant M\|\mathbf{u}\| \tag{14}$$

where

$$M=\left(\int_a^b\int_a^b [K(x,\ \xi)]^2\,dx\,d\xi\right)^{\frac{1}{2}}$$

A linear operator satisfying condition (14) is called **bounded**, and the least value of the constant M for which inequality (14) is fulfilled is called the **norm** of the operator (compare with similar definitions given for functionals in Sec. VI.1.5). Hence, an integral operator A with a kernel satisfying condition (7) is a bounded linear operator in $L_2[a, b]$. Consequently, equation (6) can be rewritten in the form $\mathbf{u}=\mathbf{Au}+\mathbf{f}$. This enables us to include the theory of Fredholm's integral equations into a more general theory of **operator equations.** We shall discuss this question in Sec. 2.11.

A Hilbert space whose elements are vector-valued functions defined in a domain belonging to a manifold of an arbitrary dimension possesses the same properties. An analogous situation takes place when the integration is performed with respect to an arbitrary nonnegative measure μ.

§ 2. FREDHOLM THEORY

In §§ 2 and 3, unless otherwise stated, we shall suppose that the kernels satisfy condition (7) and that the nonhomogeneity terms and the solutions of the equations belong to $L_2[a, b]$. But, for simplicity's sake, the reader may suppose that the functions are continuous. We consider here Fredholm's equations of the second kind whose theory is much simpler than the theory of the equations of the first kind, emphasis being laid upon one-dimensional equations. It is also supposed that the functions under consideration may assume complex values.

1. Equations with Degenerate Kernels. A kernel $K(x, \xi)$ is said to be **degenerate** if it can be represented in the form of a sum of products of functions dependent solely on x or solely on ξ:

$$K(x,\ \xi)=\sum_{j=1}^{p}\Phi_j(x)\,\Psi_j(\xi) \tag{1}$$

For example, a polynomial in x and ξ is obviously a degenerate kernel.

Without loss of generality we can suppose that the systems of functions $\Phi_j(x)$ and $\Psi_j(\xi)$ are linearly independent. Indeed, if, for instance, $\Phi_1(x)$ is expressed linearly in terms of the remaining

$\Phi_j(x)$'s, then substituting this expression into (1) and combining the terms containing the same functions Φ_j we obtain a representation of the kernel with a smaller number of summands (verify this argument!).

We shall consider equation (1.6) with an additional factor (parameter) λ in the kernel (this proves to be useful in many problems):

$$u(x)=\lambda\int_a^b K(x,\ \xi)\,u(\xi)\,d\xi+f(x)=$$

$$=\lambda\int_a^b\sum_{j=1}^{p}\Phi_j(x)\,\Psi_j(\xi)\,u(\xi)\,d\xi+f(x) \tag{2}$$

It turns out that this equation can be reduced to a system of p linear algebraic equations with p unknowns. In fact, let us suppose that for a given λ there exists a solution $u(x)$ of equation (2). If we introduce the notation

$$u_j=\int_a^b\Psi_j(x)\,u(x)\,dx=\int_a^b\Psi_j(\xi)\,u(\xi)\,d\xi \qquad (j=1,\ 2,\ \ldots,\ p) \tag{3}$$

it follows from (2) that

$$u(x)=\lambda\sum_{j=1}^{p}u_j\Phi_j(x)+f(x) \tag{4}$$

Denoting the summation index by k and substituting (4) into (3) we obtain

$$u_j-\lambda\sum_{k=1}^{p}a_{jk}u_k=f_j \qquad (j=1,\ 2,\ \ldots,\ p) \tag{5}$$

where

$$a_{jk}=\int_a^b\Psi_j(x)\,\Phi_k(x)\,dx,\quad f_j=\int_a^b\Psi_j(x)\,f(x)\,dx \qquad (j,\ k=1,\ 2,\ \ldots,\ p)$$

Conversely, it can be easily shown (let the reader do it!) that the substitution of any solution of system (5) into the right-hand side of (4) yields a solution of equation (2), equalities (3) being satisfied.

We see that formulas (3) and the reciprocal formula (4) establish a one-to-one correspondence between the solutions of integral equation (2) and the solutions of the system of linear algebraic equations (5).

As is known, the solvability of system (5) is specified by its determinant

$$D(\lambda) = \det(\delta_{jk} - \lambda a_{jk}) = \det\left(\delta_{jk} - \lambda \int_a^b \Psi_j(x)\,\Phi_k(x)\,dx\right)$$

If $D(\lambda) \neq 0$ the system has a unique solution and if

$$D(\lambda) = 0 \tag{6}$$

it has either no solution or infinitely many solutions. According to the above, equation (2) possesses the same property. $D(\lambda)$ is known as the **Fredholm determinant** of equation (2), and its zeros, that is the roots of equation (6), are called **characteristic values (numbers).** The determinant $D(\lambda)$ being a polynomial of degree not higher than p, *equation* (2) *has no more than p characteristic values.*

It also follows that *if the value of* λ *is not characteristic, equation* (2) *has exactly one solution for any function* $f(x)$. This proposition is known as **Fredholm's first theorem.** It can be formulated equivalently in the following way: *for equation* (2) *to possess exactly one solution for every function* $f(x)$ *it is necessary and sufficient that the corresponding homogeneous equation* (i.e. equation (2) in which $f(x) \equiv 0$) *have only the trivial solution* $u(x) \equiv 0$.

Solving system (5) with the aid of Cramer's formulas (e.g. see [107], Sec. VI.4) and expanding the determinants entering into the numerators in minors of the column consisting of the constant terms f_j we receive the expressions

$$u_j = \frac{1}{D(\lambda)} \sum_{k=1}^{p} D_{jk}(\lambda)\, f_k \qquad (j = 1,\ 2,\ \ldots,\ p)$$

where $D_{jk}(\lambda)$ are polynomials of degree not higher than $p-1$ (why?). Substituting these expressions into (4) we get

$$u(x) = \lambda \sum_j \frac{1}{D(\lambda)} \sum_k D_{jk}(\lambda) \int_a^b \Psi_k(\xi)\, f(\xi)\, d\xi \cdot \Phi_j(x) + f(x)$$

that is

$$u(x) = \lambda \int_a^b \Gamma(x,\ \xi,\ \lambda)\, f(\xi)\, d\xi + f(x) \tag{7}$$

where

$$\Gamma(x,\ \xi,\ \lambda) = \frac{1}{D(\lambda)} \sum_{j=1}^{p} \sum_{k=1}^{p} D_{jk}(\lambda)\, \Phi_j(x)\, \Psi_k(\xi) \tag{8}$$

The function $\Gamma(x, \xi, \lambda)$ is the **resolvent** (or **reciprocal**) **kernel** of equation (2) by means of which the sought-for solution is explicitly

expressed by formula (7). For fixed x and ξ the resolvent kernel is a rational function of the complex variable λ. Substituting (7) into (2) (to this end we change the notation of the variable of integration in the first integral (2)) and taking into account that $f(x)$ is an arbitrary function, we derive the following equation for the resolvent kernel (check it up!):

$$\Gamma(x, \xi, \lambda) = \lambda \int_a^b K(x, \eta)\, \Gamma(\eta, \xi, \lambda)\, d\eta + K(x, \xi) \tag{9}$$

In this equation λ as well as ξ are parameters.

Consider the equation

$$v(x) = \mu \int_a^b \sum_{j=1}^{p} \Phi_j^*(\xi)\, \Psi_j^*(x)\, v(\xi)\, d\xi \tag{10}$$

which is termed the **homogeneous adjoint (associated) equation** (relative to (2)). Its kernel is obtained from the original kernel by interchanging the independent variables and passing to the conjugate complex values (compare with the definition of the adjoint matrix to a given matrix). Denoting the coefficients of the type of a_{jk} and the Fredholm determinant corresponding to equation (10) by b_{jk} and $E(\mu)$, respectively, we readily obtain the formulas

$$b_{jk} = \int_a^b \Phi_j^*(x)\, \Psi_k^*(x)\, dx = a_{kj}^*,$$

$$[E(\mu)]^* = [\det(\delta_{jk} - \mu b_{jk})]^* = \det(\delta_{kj} - \mu^* b_{kj}^*) =$$
$$= \det(\delta_{jk} - \mu^* a_{jk}) = D(\mu^*),$$
$$[E_{jk}(\mu)]^* = D_{kj}(\mu^*)$$

It follows that *if λ is a characteristic value of equation (2), λ^* is a characteristic value of adjoint equation (10) and vice versa* (because equations (2) and (10) are mutually adjoint). Besides, formula (8) obviously implies (prove it!) that $(\Gamma(\xi, x, \lambda))^*$ is the resolvent kernel of equation (10) for $\mu = \lambda^*$. Taking the equation of form (9) for the latter resolvent kernel and passing to the conjugate complex values we obtain (check it up!), after interchanging x and ξ, one more equation for the resolvent kernel $\Gamma(x, \xi, \lambda)$:

$$\Gamma(x, \xi, \lambda) = \lambda \int_a^b K(\eta, \xi)\, \Gamma(x, \eta, \lambda)\, d\eta + K(x, \xi) \tag{11}$$

Now suppose that λ is a characteristic value of equation (2). The solutions of the homogeneous equation corresponding to (2) are called **eigenfunctions** of equation (2) **associated with (corresponding to)** λ (or **belonging to** λ). They constitute a linear space

whose dimension (i.e. the number of linearly independent eigenfunctions belonging to the chosen characteristic value λ) is equal to the dimension of the space of solutions of the homogeneous system of equations corresponding to (5) (why?). But this dimension is equal to $p-\text{rank}\,(\delta_{jk}-\lambda a_{jk})$ (e.g. see [107], Sec. XI.5), and therefore for $\mu=\lambda^*$ the dimension of the analogous space for equation (10) is equal to

$$p-\text{rank}\,(\delta_{jk}-\lambda^* b_{jk})=p-\text{rank}\,(\delta_{kj}-\lambda a_{jk})^*=p-\text{rank}\,(\delta_{jk}-\lambda a_{jk})$$

because any two mutually adjoint matrices are of the same rank. Thus, *the dimensions of the spaces of the eigenfunctions of equation* (2) *for the characteristic value* λ *and of adjoint equation* (10) *for* $\mu=\lambda^*$ *coincide and are finite.* This assertion is **Fredholm's second theorem.**

Let us write down equation (2) in the form $\mathbf{u}=\lambda\mathbf{A}\mathbf{u}+\mathbf{f}$ (see the end of Sec. 1.3). If λ is a characteristic value, the equation $\mathbf{u}=\lambda\mathbf{A}\mathbf{u}$, i.e. $\mathbf{A}\mathbf{u}=\frac{1}{\lambda}\mathbf{u}$, has a nontrivial solution. Consequently, $\frac{1}{\lambda}$ is an eigenvalue of the operator A, the corresponding eigenfunctions of the operator coinciding with the eigenfunctions of the integral equation. It is also clear that the reciprocal of any nonzero eigenvalue of the operator A is a characteristic value of equation (2).

It should also be noted that, by § IV.3, the dimension of the space of the eigenfunctions corresponding to a chosen characteristic value λ is either equal to or less than the multiplicity of λ as root of equation (6).

Finally, consider nonhomogeneous equation (2) when λ is a characteristic value. For at least one solution of equation (2) (i.e. of system (5)) to exist, the function f must satisfy some additional conditions. As was shown in Sec. IV.1.4, for the existence of at least one solution of a system of linear algebraic equations it is necessary and sufficient that the column of right-hand members be orthogonal to all the solutions of the associated homogeneous system of equations. For system (5) this means that the equalities

$$\sum_{j=1}^{p} f_j v_j^* = 0 \tag{12}$$

should be fulfilled for any solution v_j of the system of equations

$$v_j-\lambda^*\sum_{k=1}^{p} b_{jk}v_k=0 \qquad (j=1,\ 2,\ \ldots,\ p)$$

Equalities (12) can be rewritten in the form

$$\sum_{j=1}^{p}\left(\int_a^b \Psi_j(x)\,f(x)\,dx\right)\left(\int_a^b \Phi_j^*(x)\,v(x)\,dx\right)^*=0 \tag{13}$$

where $v(x)$ is the corresponding solution of equation (10) for $\mu = \lambda^*$. Replacing the variable of integration in the second integral by ξ we can write the product of the integrals in (13) as a double integral:

$$\int_a^b dx \int_a^b \sum_{j=1}^{p} \Psi_j(x)\Phi_j(\xi) f(x) v^*(\xi)\, d\xi = 0$$

Now using relation (10) and cancelling by λ we finally obtain

$$\int_a^b f(x) v^*(x)\, dx = 0 \tag{14}$$

Thus, *if λ is a characteristic value, then for at least one solution of equation* (2) *to exist it is necessary and sufficient that the function f be orthogonal to all the solutions of the homogeneous adjoint equation* (10) *for* $\mu = \lambda^*$. This is **Fredholm's third theorem.** If this orthogonality condition holds, the general solution of equation (2) is equal to the sum of any particular solution of the equation and the general solution of the corresponding homogeneous equation, the rule being common for nonhomogeneous linear equations of any nature.

In the formulation of Fredholm's third theorem we can omit the condition that λ is a characteristic value. Then the second and third theorems embrace the first one as a particular case (why?).

Let us deduce a formula representing the solution of equation (2) in terms of the resolvent kernel for the important special case when the value $\lambda = \lambda_0$ is a simple zero of the determinant $D(\lambda)$. In this case, by (8), we have the relation

$$\Gamma(x, \xi, \lambda) = \frac{\Gamma_0(x, \xi)}{\lambda - \lambda_0} + \Gamma_1(x, \xi, \lambda) \tag{15}$$

where the function Γ_1 is analytic for $\lambda = \lambda_0$. Substituting (15) into (9) we receive

$$\Gamma_0(x, \xi) = \lambda_0 \int_a^b K(x, \eta)\, \Gamma_0(\eta, \xi)\, d\eta$$

whence it follows that for any fixed ξ the function $\Gamma_0(x, \xi)$ must be proportional to the only linearly independent eigenfunction $u_0(x)$ of the kernel K corresponding to the value λ_0. Applying the same argument to (11) we conclude that $\Gamma_0(x, \xi) = C_0 u_0(x)\, [v_0(\xi)]^*$ where $v_0(x)$ is the only eigenfunction of the adjoint kernel for $\lambda = \lambda_0^*$.

Substituting expression (15) into (7) we get

$$u(x) = \frac{\lambda}{\lambda - \lambda_0} \int_a^b C_0 u_0(x)\, [v_0(\xi)]^* f(\xi)\, d\xi + \lambda \int_a^b \Gamma_1(x, \xi, \lambda) f(\xi)\, d\xi + f(x)$$

Now letting $\lambda \to \lambda_0$ we again see that the function $f(x)$ must be orthogonal to the eigenfunction $v_0(x)$. If this condition is fulfilled then for $\lambda = \lambda_0$ we obtain

$$u(x) = \lambda_0 \int_a^b \Gamma_1(x, \xi, \lambda_0) f(\xi)\, d\xi + f(x) + Cu_0(x)$$

where C is an arbitrary constant. This is the desired representation. In the case of a zero of an arbitrary multiplicity the representation is written in the same form if Γ_1 is understood as the regular part of the expansion of the resolvent kernel into Laurent's series at the point λ_0.

It should be noted that in the above considerations we did not make use of the fact that the systems of functions Φ_j and Ψ_j are linearly independent. In the case of linear dependence the number of the functions can be reduced, which obviously simplifies the calculations.

2. General Case. Now let us consider Fredholm's general equation of the second kind

$$u(x) = \lambda \int_a^b K(x, \xi)\, u(\xi)\, d\xi + f(x) \tag{16}$$

with the parameter λ entering as a factor in the kernel.

The kernel K can be approximated with degenerate kernels to an arbitrary accuracy. Such an approximation can be performed in many ways, for instance, by means of Fourier's double series in trigonometric functions (e.g. see [107], Sec. XVII.30), by using some other complete orthogonal system of functions, by constructing an appropriate polynomial in x and ξ, and the like. The approximation can be performed in the sense of the space C (i.e. with respect to uniform deviation of functions) if the kernel is continuous and the limits of integration a and b are finite, and in the sense of L_2 if it is discontinuous (see [71]).

Replacing the kernel by a degenerate one we obtain an equation to which all the results of Sec. 1 are applicable. Therefore, *the Fredholm theorems hold for general equation* (16) *whose kernel satisfies condition* (1.7). (In the comprehensive courses on integral equations mentioned above the validity of such an approach is rigorously justified.) The only essential distinction from Sec. 1 is that in the general case Fredholm's determinant $D(\lambda)$ is no longer a polynomial. The greater the accuracy of the approximation of the kernel by degenerate kernels, the greater the number of summands, and in the limit $D(\lambda)$ becomes "a polynomial of an infinite degree", that is an entire analytic function. Therefore, equation (16) may have an infinite number of characteristic values (but, as we shall see later, it may also have a finite

number of characteristic values or none). But $D(0)=1$, and therefore $D(\lambda) \not\equiv 0$. Consequently, according to the theory of analytic functions (see Sec. II.2.6), we conclude that *if there is an infinitude of characteristic values they can be arranged as an infinitely large sequence* $\lambda_1, \lambda_2, \lambda_3, \ldots$. This proposition is **Fredholm's fourth theorem.**

Accordingly, in the general case the resolvent kernel $\Gamma(x, \xi, \lambda)$ is equal to the ratio of two entire analytic functions and therefore is no longer a rational function but a meromorphic one (see Sec. II.3.12).

In the general case the homogeneous adjoint equation has the form

$$v(x)=\mu \int_a^b L(x, \xi)\, v(\xi)\, d\xi \text{ where } L(x, \xi)=[K(\xi, x)]^* \qquad (17)$$

(instead of (10)). There is a simple relation connecting the original kernel K and the adjoint kernel L: for any functions $\varphi(x)$ and $\psi(x)$ we have

$$\int_a^b \left[\int_a^b K(x, \xi)\, \varphi(\xi)\, d\xi\right] \psi^*(x)\, dx = \int_a^b \varphi(x) \left[\int_a^b L(x, \xi)\, \psi(\xi)\, d\xi\right]^* dx \qquad (18)$$

(check it up!).

If the original equation contains no parameter (like equation (1.6)) then, to apply the first three Fredholm theorems, we must put $\lambda=\mu=1$. Instead of saying "if 1 is not a characteristic value" we can say "if a homogeneous equation associated with the given equation has only a trivial solution". In this case we also say that **Fredholm's first alternative** holds; if otherwise, we say that **Fredholm's second alternative** takes place.

The above theoretical approach can be used as a method for numerical solution of equation (16) for concrete λ as well as a method for calculating characteristic values and eigenfunctions of this equation.

As an example, let us consider the equation

$$u(x)=\lambda \int_0^1 \sin(x\xi)\, u(\xi)\, d\xi + 1$$

We shall find its characteristic values and calculate the solution for $\lambda=\frac{1}{2}$. To this end use the expansion of the sine into Maclaurin's series:

$$\sin(x\xi)=x\xi-\frac{x^3\xi^3}{6}+\frac{x^5\xi^5}{120}-\ldots \qquad (19)$$

Taking the first term we obtain the approximate equation

$$u(x)=\lambda\int_0^1 x\xi u(\xi)\,d\xi+1$$

According to Sec. 1, we now put

$$u_1=\int_0^1 \xi u(\xi)\,d\xi \qquad (20)$$

which yields

$$u(x)=\lambda u_1 x+1$$

Substituting this expression into (20) we get

$$u_1=\int_0^1 \xi(\lambda u_1\xi+1)\,d\xi=\lambda\frac{u_1}{3}+\frac{1}{2}$$

whence we find the determinant $D^{(1)}(\lambda)$:

$$D^{(1)}(\lambda)=1-\frac{\lambda}{3}=1-0.3333\lambda$$

Hence, there is only one characteristic value $\lambda_1^{(1)}=3$. Accordingly, for $\lambda=\frac{1}{2}$ we obtain the solution

$$u^{(1)}(x)=1+0.3x \qquad (0\leqslant x\leqslant 1)$$

Taking the first two terms of series (19) we receive (check it up!)

$$u(x)=\lambda\int_0^1\left(x\xi-\frac{x^3\xi^3}{6}\right)u(\xi)\,d\xi+1,$$

$$u_1=\int_0^1 \xi u(\xi)\,d\xi,\quad u_2=\int_0^1 \xi^3 u(\xi)\,d\xi,\quad u(x)=\lambda u_1 x-\lambda u_2\frac{x^3}{6}+1,$$

$$\left.\begin{aligned} u_1-\lambda\left(\frac{u_1}{3}-\frac{u_2}{30}\right)&=\frac{1}{2}\\ u_2-\lambda\left(\frac{u_1}{5}-\frac{u_2}{42}\right)&=\frac{1}{4}\end{aligned}\right\} \qquad (21)$$

$$D^{(2)}(\lambda)=\begin{vmatrix} 1-\frac{\lambda}{3} & \frac{\lambda}{30}\\ -\frac{\lambda}{5} & 1+\frac{\lambda}{42}\end{vmatrix}=1-\frac{13}{42}\lambda-\frac{2}{1575}\lambda^2=1-0.3095\lambda-0.0013\lambda^2,$$

$$\lambda_1^{(2)}=3.16,\quad \lambda_2^{(2)}=-247$$

$$u^2(x)=1+0.297x-0.025x^3 \qquad (0\leqslant x\leqslant 1)$$

Here, when passing from $p=1$ to $p=2$, we obviously see a tendency to convergence in the expressions of $D(\lambda)$, λ_1 and $u(x)$.

We suggest that the reader consider the case $p=3$, that is take the first three terms of series (19), and find the corresponding approximations. When solving a system of form (21) for a relatively small λ we can use the iteration method (e.g. see [107], Sec. VI.5). Besides, when solving algebraic equations whose roots considerably differ in their absolute values it is convenient to use the well-known algebraic relations connecting the roots (e.g. see [107], formula (VIII.33)). For instance, the equation $D^{(2)}(\lambda)=0$ can be solved in the following way: we first rewrite it in the form $\lambda=\frac{(1-0.0013\lambda^2)}{0.3095}$, then, beginning with $\lambda_1^{(1)}=3$, we find $\lambda_1^{(2)}$ by means of iterations and finally apply the formula for the sum of the roots of a quadratic equation.

Here we mention one more method of numerical solution of equation (16) based on formula (18) which resembles the method of moments (e.g. see [107], Sec. XV.29). Replacing in formula (18) the function $\varphi(x)$ by the sought-for solution $u(x)$ of equation (16) we obtain the equality

$$\int_a^b u(x)\left[\psi(x)-\lambda^*\int_a^b L(x,\xi)\,\psi(\xi)\,d\xi\right]^* dx=\int_a^b f(x)\,\psi^*(x)\,dx$$

which is fulfilled for any function $\psi(x)$. Let us choose a system of functions $\psi_1(x)$, $\psi_2(x)$, ..., $\psi_n(x)$ (for this purpose we usually take the first n terms of a complete system of functions) and substitute each of them for $\psi(x)$. This results in

$$\int_a^b u(x)\,g_j^*(x)\,dx=\int_a^b f(x)\,\psi_j^*(x)\,dx \qquad (j=1,\ 2,\ \ldots,\ n) \qquad (22)$$

where

$$g_j(x)=\psi_j(x)-\lambda^*\int_a^b L(x,\ \xi)\,\psi_j(\xi)\,d\xi$$

Thus, the sought-for function $u(x)$ satisfies n linear relations and therefore it is natural to take it in the form of a linear combination

$$u(x)=\sum_{k=1}^{n} C_k h_k(x) \qquad (23)$$

of a system of functions $h_k(x)$ with n arbitrary coefficients. The substitution of (23) into (22) leads to a system of n linear algebraic equations from which we can find these coefficients and, consequently, approximate solution (23) of equation (16). As a rule, as $h_k(x)$ we take the first few terms of a complete system of functions. If λ is not a characteristic value we can put $h_k(x)=g_k(x)$

because in this case it can readily be verified that the system of functions g_k is complete. (Prove this property on the basis of the following equivalent definition of completeness: *a system of functions belonging to* $L_2[a, b]$ *is complete if there is no nonzero function orthogonal to all the functions of the system.*) The system of functions g_j can be orthogonalized, and then the system of equations for C_k becomes diagonal.

The general results given above can be directly extended to the equations of form (1.9) with unknown functions of several variables, to equations (1.10) and to the systems of m equations with m unknown functions. Such systems can be written in form (16) or (1.9) with K interpreted as an $m \times m$ matrix and u and f as vectors. In this case the orthogonality condition similar to (14) takes the form $\int f^* v dG = 0$.

3. Applying Infinite Systems of Algebraic Equations. In the example of Sec. 2 we could use complete series (19) with an infinite number of terms. In some other cases the kernel can be expanded in a series with respect to degenerate kernels:

$$K(x, \xi) = \sum_{j=1}^{\infty} \Phi_j(x) \Psi_j(\xi)$$

This enables us, at least theoretically, to construct an exact solution by using the method of Sec. 1 with the upper limit of summation p replaced by ∞, i.e. in the form of the sum of an infinite series. Then, instead of (5), we obtain the infinite system

$$u_j - \lambda \sum_{k=1}^{\infty} a_{jk} u_k = f_j \qquad (i = 1, 2, 3, \ldots) \tag{24}$$

with an infinite number of unknowns. In practice we usually solve such equations by **truncation**, that is by passing to system (5) for an appropriately chosen finite number p of terms. Hence, we arrive at the same system which is obtained if we at the very beginning approximately replace the kernel K by a degenerate one. But in some cases we may succeed in obtaining an exact solution of complete infinite system (24) itself or in deriving some useful consequences from it.

For instance, such an exact solution can be found if the systems of functions $\Phi_j(x)$ and $\Psi_j(x)$ are biorthogonal, i.e.

$$a_{jk} = \int_a^b \Psi_j(x) \Phi_k(x) dx = 0 \qquad (j \neq k)$$

Indeed, in this case nondiagonal terms of system (24) vanish, and

its exact solution is written in the form

$$u_j = \left(1 - \lambda \int_a^b \Psi_j(x)\,\Phi_j(x)\,dx\right)^{-1} f_j \qquad (j = 1,\ 2,\ 3,\ \ldots)$$

The characteristic values are equal to

$$a_{nn}^{-1} = \left(\int_a^b \Psi_n(x)\,\Phi_n(x)\,dx\right)^{-1} \qquad (n = 1,\ 2,\ 3,\ \ldots)$$

(if $a_{nn} \neq 0$), and the corresponding eigenfunctions coincide with the functions $\Phi_n(x)$.

There is another important particular case when the infinite matrix (a_{jk}) is triangular, i.e. such that either all $a_{jk} = 0$ for $j > k$ (an upper triangular matrix) or $a_{jk} = 0$ for $j < k$ (a lower triangular matrix). In both cases, equating the determinant to zero we find the characteristic values $\lambda_n = a_{nn}^{-1}$ $(n = 1, 2, 3, \ldots)$. For the upper triangular matrix it is easy to find the corresponding eigenfunctions because putting $u_{n+1} = u_{n+2} = \ldots = 0$ in homogeneous system (24) we arrive at a homogeneous system of n equations with n unknowns and zero determinant. Therefore, the eigenfunctions are finite linear combinations of the functions $\Phi_j(x)$. (For instance, prove that $\Phi_1(x)$ is an eigenfunction corresponding to the value λ_1, $\Phi_2(x) - \frac{a_{12}}{a_{11} - a_{22}}\Phi_1(x)$ corresponds to λ_2, etc.; here, it is of course supposed that all the denominators we deal with are different from zero.)

If $a_{jk} = 0$ for all $j > k$ and all $j < k - 1$ we have a particularly simple case for which homogeneous system (24) takes the form

$$(1 - \lambda a_{jj})\,u_j - \lambda a_{j,\,j+1} u_{j+1} = 0 \qquad (j = 1,\ 2,\ 3,\ \ldots)$$

whence

$$u_1 = u_1,\quad u_2 = \frac{1 - \lambda a_{11}}{\lambda a_{12}}\,u_1,\quad u_3 = \frac{1 - \lambda a_{22}}{\lambda a_{23}}\,u_2 = \frac{(1 - \lambda a_{22})\,(1 - \lambda a_{11})}{(\lambda a_{23})\,(\lambda a_{12})}\,u_1$$

and, generally,

$$u_m = \left(\prod_{j=1}^{m-1} \frac{1 - \lambda a_{jj}}{\lambda a_{j,\,j+1}}\right) u_1 \qquad (m - 1,\ 2,\ 3,\ \ldots)$$

(for $m = 1$ the last expression is understood as being equal to unity). Hence, it follows that for $\lambda = a_{nn}^{-1}$ we again have $u_{n+1} = u_{n+2} = \ldots = 0$ and thus arrive at explicit expressions for $u_1, u_2, \ldots, u_n$. It can be proved that in the general case all the other values of λ lead to divergent series for $u(x)$, that is do not yield eigenfunctions.

For the lower triangular matrix (a_{jk}) the solution of nonhomogeneous system (24) can also be easily found (how?). However, the

expressions of the eigenfunctions are more complicated in this case than for an upper triangular matrix. For example, substituting $\lambda=\lambda_1=a_{11}^{-1}$ into homogeneous system (24) we obtain, in succession,

$$u_2=\frac{a_{21}}{a_{11}-a_{22}}u_1$$

$$u_3=\frac{a_{31}u_1+a_{32}u_2}{a_{11}-a_{33}}=\left[\frac{a_{31}}{a_{11}-a_{33}}+\frac{a_{21}a_{32}}{(a_{11}-a_{22})(a_{11}-a_{33})}\right]u_1$$

etc. Substituting these expressions into formula (4) for $p=\infty$ and $f\equiv 0$ we receive the first eigenfunction in the form of a sum of a series. When calculating the eigenfunction corresponding to the value $\lambda=\lambda_n=a_{nn}^{-1}$ we must put $u_1=u_2=\ldots=u_{n-1}=0$ and, then, using an analogous procedure, express the remaining coefficients u_j in terms of u_n.

Infinite systems of equations of form (24) are dealt with in various divisions of applied mathematics but here we are not going to treat the general properties of such systems (e.g. see [58], § I.2). In problem solving practice such systems are often rewritten in the form

$$u_j=\lambda\sum_{k=1}^{\infty}a_{jk}u_k+f_j \qquad (j=1,\ 2,\ 3,\ \ldots) \tag{25}$$

after which it is natural to apply the ordinary iteration process which converges if all the coefficients λa_{jk} are sufficiently small. If the coefficients with small j and k are not small we can resolve a number (say n) of the first equations (25) with respect to $u_1,\ u_2,\ \ldots,\ u_n$, add the remaining equations (25) to the resulting formulas and again apply the iteration method.

In the case when $a_{jk}=0$ for $|j-k|>1$ we have another interesting class of systems (24) which is encountered in some applications. Here we shall only consider homogeneous systems, i.e. those having the form

$$\left.\begin{aligned}(1-\lambda a_{11})\,u_1-\lambda a_{12}u_2&=0,\\ -\lambda a_{21}u_1+(1-\lambda a_{22})\,u_2-\lambda a_{23}u_3&=0,\\ -\lambda a_{32}u_2+(1-\lambda a_{33})\,u_3-\lambda a_{34}u_4&=0,\\ \ldots\ldots\ldots\ldots\ldots\ldots&\end{aligned}\right\}$$

Putting $\frac{u_{k+1}}{u_k}=U_k$, $\frac{1}{a_{k,k-1}}=p_k$, $\frac{a_{kk}}{a_{k,k-1}}=q_k$ and $\frac{a_{k,k+1}}{a_{k,k-1}}=r_k$ we obtain (check it up!)

$$U_1=\frac{\lambda^{-1}-a_{11}}{a_{12}},\quad 1-(p_k\lambda^{-1}-q_k)\,U_{k-1}+r_kU_{k-1}U_k=0$$
$$(k=2,\ 3,\ 4,\ \ldots) \tag{26}$$

It follows that

$$U_{k-1}=\frac{1}{p_k\lambda^{-1}-q_k-r_kU_k} \tag{27}$$

For $k=2$ this yields

$$U_1=\frac{1}{p_2\lambda^{-1}-q_2-r_2U_2} \tag{28}$$

Next we substitute $k=3$ into (27) and the result obtained into (28), which gives

$$U_1=\cfrac{1}{p_2\lambda^{-1}-q_2-\cfrac{r_2}{p_3\lambda^{-1}-q_3-r_3U_3}} \tag{29}$$

Proceeding in this way we substitute $k=4$ into (27) and the resultant expression into (29) and so on. We thus arrive at a so-called **continued fraction** which is conditionally written as

$$U_1=\cfrac{1}{p_2\lambda^{-1}-q_2-}\cfrac{r_2}{p_3\lambda^{-1}-q_3-}\cfrac{r_3}{p_4\lambda^{-1}-q_4-}\cdots \tag{30}$$

Continued fractions have many interesting properties and various applications which will not be discussed here. In practical calculations such fractions are truncated, that is they are replaced by finite fractions. This is equivalent to putting $U_n=0$ for some n in recurrence formula (26) and finding, in succession, the quantities U_{n-1}, U_{n-2}, ..., U_1 from it.

Formula (30) and the first formula (26) yield an equation for the characteristic values λ. If this equation is solved by means of an approximate method, then from (26) it is easy to find, in succession, U_1, U_2, U_3, ... and thus determine the corresponding eigenfunction.

In conclusion it should be noted that system (24) may have only a finite number of coefficients a_{jk} different from zero. From the theoretical point of view this case is the simplest.

4. Applying Numerical Integration. There is another approach to Fredholm's theorems which can also serve as a basis for numerical methods. Let us take integral equation (16) and apply numerical integration to the integral involved (e.g. see [107], Sec. XIV.13). Every formula of numerical integration has the form

$$\int_a^b \varphi(x)\,dx\approx\sum_{k=1}^{n}H_k\varphi(x_k) \tag{31}$$

where x_k are nodes and H_k are certain coefficients. Using formula (31), substituting a nodal value x_j into (16) and replacing the sign $\approx$ by $=$ we obtain

$$u_j=\lambda\sum_{k=1}^{n}H_kK(x_j,\ x_k)\,u_k+f_j \qquad (j=1,\ 2,\ \ldots,\ n) \tag{32}$$

where

$$u_j = u(x_j), \qquad f_j = f(x_j)$$

For instance, we can put

$$x_k = a + kh - \frac{h}{2}, \quad H_k = h, \quad \text{where } h = \frac{b-a}{n} \tag{33}$$

then equalities (32) take the form

$$u_j = \lambda h \sum_{k=1}^{n} K(x_j, x_k) u_k + f_j \qquad (j = 1, 2, \ldots, n) \tag{34}$$

In other words, we have taken advantage of the fact that an integral can be approximately regarded as a sum of a great number of small summands. Thus, we have directly come to a system of n linear algebraic equations with n unknowns. For such systems Fredholm's theorems are equivalent to the conditions guaranteeing their solvability and therefore in the limit, for $n \to \infty$, these theorems are extended to equations (16).

This method can also be given a rigorous mathematical justification. Nevertheless, it was preferable to begin with degenerate kernels not only because in many simple problems their application is most effective but also because an equation with a degenerate kernel is an integral equation while system of equations (32) is essentially connected with an additional passage to the limit.

For practical purposes it is desirable to choose a formula of type (31) which provides a greater accuracy for the same amount of calculations. Various formulas for numerical integration can be found in courses on approximate calculations and special monographs devoted to numerical integration, e.g. see [74]. Here we shall only dwell on **Gauss' method** which is one of the most effective methods.

First of all we pass from the integration over the interval $[a, b]$ to the standard interval $[-1, 1]$ by means of a linear change of the variable of integration: $x = a + (b-a)\frac{x_1+1}{2}$. In what follows we shall assume that this change has already been made. Let us try, for a chosen n, to determine the coefficients H_k and the nodes x_k in such a way that the formula

$$\int_{-1}^{1} \varphi(x)\, dx \approx \sum_{k=1}^{n} H_k \varphi(x_k) \tag{35}$$

becomes exact for all polynomials of degree not higher than p where p is as great as possible. For this purpose it is sufficient that formula (35) be exact for the functions $\varphi(x) = 1, x, x^2, \ldots, x^p$ (why?). Substituting these functions into (35) we obtain $p+1$

equations

$$\int_{-1}^{1} x^s dx = \sum_{k=1}^{n} H_k x_k^s \qquad (s=0, 1, 2, \ldots, p) \tag{36}$$

which must be satisfied. Since for arbitrarily chosen coefficients and nodes there are $2n$ degrees of freedom, we have every reason to expect that in the best case $p=2n-1$. Gauss' formula is a realization of this heuristic argument.

For our further considerations we need Legendre's polynomials (e.g. see [107], Sec. XVII.20). It can be proved that all the zeros of the Legendre polynomial $P_n(x)$ are real and simple and lie in the interval $-1<x<1$ (we do not present the proof here but for a number of the first polynomials this assertion can be verified directly). It is these zeros that are taken as the nodes x_k (the expedience of this choice will be elucidated below). After the nodes have been chosen, we put $s=0, 1, \ldots, n-1$ in equalities (36); this results, for the coefficients H_k, in a system of n equations with n unknowns from which they are found.

Let us show that for the nodes and coefficients thus determined equalities (36) are fulfilled for $s=0, 1, \ldots, 2n-1$. Indeed, for $0\leqslant s\leqslant n-1$ this is true by the construction of the coefficients, and therefore equality (35) is exact for all polynomials of degree not higher than $n-1$. Now let $n\leqslant s\leqslant 2n-1$. Dividing x^s by $P_n(x)$ we obtain a polynomial $Q_{s-n}(x)$ of degree $s-n$ as the quotient and a polynomial $R_{<n}(x)$ of degree less than n as the remainder. Consequently, $x^s = Q_{s-n}(x)\,P_n(x)+R_{<n}(x)$, and hence equality (36) which must be proved takes the form

$$\int_{-1}^{1} Q_{s-n}(x)\,P_n(x)\,dx+\int_{-1}^{1} R_{<n}(x)\,dx=$$

$$=\sum_{k=1}^{n} H_k Q_{s-n}(x_k)\,P_n(x_k)+\sum_{k=1}^{n} H_k R_{<n}(x_k) \tag{37}$$

But, since $s-n<n$, the polynomial $Q_{s-n}(x)$ can be represented as a linear combination of Legendre's polynomials of degree less than n (why?). It then follows that, by the orthogonality of Legendre's polynomials, the first integral in (37) is equal to zero (this is the reason why in this method we use Legendre's polynomials). According to the choice of the nodes x_k, the first sum on the right-hand side of (37) is equal to zero. Discarding the zero terms in (37) we see that now it only remains to prove an (exact) equality of type (35) for a polynomial of degree less than n, which has been done above.

The nodal values x_k and the coefficients H_k can be found in reference books. They are also contained in special subroutines

for numerical integration by Gauss' method on electronic digital computers. As an example we give here these values for small n:

$$n=2: \quad x_{1,2}=\pm 0.57735, \quad H_{1,2}=1,$$

$$n=3: \quad x_{1,3}=\pm 0.77460, \quad x_2=0, \quad H_{1,3}=\frac{5}{9}, \quad H_2=\frac{8}{9},$$

$$n=4: \quad x_{1,4}=\pm 0.86114, \quad x_{2,3}=\pm 0.33998,$$
$$H_{1,4}=0.34785, \quad H_{2,3}=0.65215,$$

$$n=5: \quad x_{1,5}=\pm 0.90618, \quad x_{2,4}=\pm 0.53847, \quad x_3=0,$$
$$H_{1,5}=0.23693, \quad H_{2,4}=0.47863, \quad H_3=0.56889$$

Techniques similar to Gauss' method are used in the **Newton-Cotes** and **Chebyshev's integration methods.** In the former method the nodes $x_k=-1+2\frac{k-1}{n-1}$ $(k=1, 2, \ldots, n)$ are fixed and the coefficients H_k are selected so that formula (35) becomes exact for all polynomials of degree not higher than $n-1$. In the latter method all the coefficients H_k are taken equal and the nodes x_k are chosen in such a way that formula (35) is exact for all polynomials of degree less than n. The corresponding values of the coefficients and nodes can be found in reference books. It is interesting to note that in Chebyshev's method only the values $n \leqslant 7$ and $n=9$ are admissible.

It should also be noted that when finding the coefficients H_k from system of equations (36) we encounter the determinant

$$D=\begin{vmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \\ x_1^2 & x_2^2 & \ldots & x_n^2 \\ \ldots & \ldots & \ldots & \ldots \\ x_1^{n-1} & x_2^{n-1} & \ldots & x_n^{n-1} \end{vmatrix} \tag{38}$$

which occurs (for arbitrary values of the numbers x_k) in various mathematical problems and is called the **Vandermonde determinant.** Its value is expressed by the formula

$$D=(x_2-x_1)(x_3-x_1)\ldots(x_n-x_1)(x_3-x_2)\ldots$$
$$\ldots(x_n-x_2)\ldots(x_n-x_{n-1}) \tag{39}$$

To derive formula (39) let us regard $x_1, x_2, \ldots, x_n$ as independent variables. Then D is a multilinear form (i.e. a homogeneous polynomial) of degree $1+2+\ldots+(n-1)=\frac{(n-1)n}{2}$. If we subtract the jth column of determinant (38) from its ith column, all the elements of the resultant column will be divisible by x_i-x_j (why?). Hence, D is divisible by every difference in the brackets entering

into the right-hand side of (39). These differences having no common factor, D is divisible by their product. Since this product is a polynomial of the same degree as D (why?), it may only differ from D in a constant factor. But the product of the elements of the principal diagonal of a determinant always enters into the expansion of the determinant with the plus sign, and in this case it is just equal to the product of the first terms of all the brackets in the right-hand side of (39). Therefore, the above constant factor can only be equal to unity, i.e. formula (39) has been proved.

In particular, this formula implies that if all the numbers x_k are different from each other the Vandermonde determinant is nonzero, which is essential for the possibility of determining the coefficients H_k.

Let us also consider equations (6) with a singularity of the polar type (Sec. 1.2) for $x=\xi$. Rapid and irregular variations of the kernel in the vicinity of the diagonal $x=\xi$ (even if not the kernel itself but its derivative has discontinuities) are an additional source of errors in the approximate solution. If the kernel approaches infinity for $x=\xi$, special measures should be taken for the diagonal terms in the formula of numerical integration not to become too significant since they are not in fact dominant because of the convergence of the integrals.

There are various methods used in practical calculations in such cases. For instance, if $K(x_k, x_k)=\pm\infty$, we can use, instead of $K(x_k, x_k)$, the value $\frac{[K(x_k, x_k+h)+K(x_k, x_k-h)]}{2}$ where h is of the order of the step Δx_k. We can also transform equation (16) to the form

$$\left[1-\lambda\int_a^b K(x, \xi)\,d\xi\right]u(x)=\lambda\int_a^b K(x, \xi)\,[u(\xi)-u(x)]\,d\xi+f(x)$$

before applying formulas of numerical integration and then replace by zero, in the right-hand side, the values of the integrand taken at the points $x=\xi$ (why?).

One can also resort to the method put forward in 1929 by the Soviet mathematicians N.M. Krylov (1879-1955) and N.N. Bogolyubov (born 1909) in which discrete values of the unknown function and the averaged values of the kernel are used. For instance, let us take the nodes and the coefficients specified by formula (33). Then we can approximately write

$$\int_a^b K(x_j, \xi)\,u(\xi)\,d\xi\approx\sum_{k=1}^{n}\left[\int_{a+(k-1)h}^{a+kh} K(x_j, \xi)\,d\xi\right]u(x_k)$$

When calculating the coefficients we integrate the kernel and

therefore they do not turn into infinity. (By the way, this method can also be applied to singular kernels; see Sec. 1.2.) For simpler kernels this procedure may lead to the exact values of the coefficients. In more complicated cases it is possible to use some more sophisticated methods of numerical integration applicable to functions with singularities (e. g. see [107], Sec. XIV.16).

5. Equations with Small Kernels. As is known (e.g. see [107], Secs. VI.5 and XVII.18), a system of linear algebraic equations of the form

$$x_j = \sum_{k=1}^{n} a_{jk} x_k + c_j \qquad (j = 1, 2, \ldots, n)$$

with sufficiently small coefficients a_{jk} can be solved with the aid of the simplest iteration scheme. Namely, a zeroth approximation is substituted into the right-hand sides, which results in the first approximation; the latter is again substituted into the right-hand sides, and so on. The iterations converge, that is in the limit (or, practically, after a certain number of iterations) we obtain the solution. Equation (16) can be approximately written in the form of linear algebraic system (32), which indicates that *the iterations converge for equation* (16) *as well if* $|\lambda|$ *is sufficiently small*, i.e. the limit they tend to is equal to exact solution.

Let us describe the iteration process in more detail. We shall take the zero approximation

$$u_0(x) = f(x)$$

(which is obtained if we put $\lambda = 0$ in (16)) though, theoretically, we can begin with any $u_0(x)$. (For practical purposes, to accelerate the convergence, it is desirable to choose $u_0(x)$ as close as possible to the exact solution if there is any information about it.) Substituting $u_0(x)$ into the right-hand side of (16) we obtain the first approximation

$$u_1(x) = f(x) + \lambda \int_a^b K(x, \xi) f(\xi)\, d\xi \tag{40}$$

Replacing in (40) the variable of integration ξ by η and the independent variable x by ξ we obtain $u_1(\xi)$ and then compute $u_2(x)$:

$$u_2(x) = f(x) + \lambda \int_a^b K(x, \xi) \left[f(\xi) + \lambda \int_a^b K(\xi, \eta) f(\eta)\, d\eta \right] d\xi =$$

$$= f(x) + \lambda \int_a^b K(x, \xi) f(\xi)\, d\xi + \lambda^2 \int_a^b K(x, \xi)\, d\xi \int_a^b K(\xi, \eta) f(\eta)\, d\eta \tag{41}$$

Reversing the order of integration in the last integral and changing the notation of the variables of integration we obtain

$$\int_a^b f(\eta)\,d\eta \int_a^b K(x, \xi) K(\xi, \eta)\,d\xi =$$

$$= \int_a^b \left(\int_a^b K(x, \eta) K(\eta, \xi)\,d\eta \right) f(\xi)\,d\xi = \int_a^b K_2(x, \xi) f(\xi)\,d\xi$$

where

$$K_2(x, \xi) = \int_a^b K(x, \eta) K(\eta, \xi)\,d\eta$$

is the **iterated kernel**. The substitution of this expression into (41) results in

$$u_2(x) = f(x) + \lambda \int_a^b K(x, \xi) f(\xi)\,d\xi + \lambda^2 \int_a^b K_2(x, \xi) f(\xi)\,d\xi$$

The repetition of the process leads to the general expression

$$u_n(x) = f(x) + \sum_{k=1}^{n} \lambda^k \int_a^b K_k(x, \xi) f(\xi)\,d\xi, \quad n = 1, 2, \ldots \qquad (42)$$

(where $K_1(x, \xi) = K(x, \xi)$), and the subsequent iterated kernels are determined by the formula

$$K_{s+1}(x, \xi) = \int_a^b K_s(x, \eta) K(\eta, \xi)\,d\eta, \qquad s = 1, 2, \ldots \qquad (43)$$

whose right member is the so-called **convolution** of the kernels $K_s(x, \xi)$ and $K(x, \xi)$. This operation is completely analogous to multiplication of matrices (e.g. see [107], Sec. XI.2); moreover, it turns into matrix multiplication if, when calculating integral (43), we apply the formula of numerical integration (31) with the nodes and coefficients determined by formula (33) and replace each kernel $K_\nu(x, \xi)$ $(\nu = s, s+1)$ by the set of values $hK_\nu(x_j, x_k)$ (check it up!). That is why the convolution of kernels possesses the properties of the multiplication of matrices.

From (43) it is clear (and it can be proved more rigorously) that the rate of increase of the sequence of kernels K_k is similar to that of a geometric progression. This means that if $|\lambda|$ is less than the reciprocal of the common ratio of this progression, then, for increasing n, the series on the right-hand side of (42) and, together with it, the whole iteration process are convergent, that is

in the limit we obtain the exact solution

$$u(x) = f(x) + \sum_{k=1}^{\infty} \lambda^k \int_a^b K_k(x, \xi) f(\xi)\, d\xi = f(x) + \lambda \int_a^b \Gamma(x, \xi, \lambda) f(\xi)\, d\xi \tag{44}$$

of equation (16) where

$$\Gamma(x, \xi, \lambda) = \sum_{k=1}^{\infty} \lambda^{k-1} K_k(x, \xi) \tag{45}$$

This is the **Neumann series** representing the resolvent kernel (see (7)) of equation (16) for sufficiently small $|\lambda|$. It follows from the theory of analytic functions (Sec. II.2.5) that the radius of convergence of this series is equal to the least of the moduli of the singular points of the function Γ (considered as an analytic function of λ). It can also be shown that this radius is equal to the least of the moduli of the characteristic values of equation (16). For large $|\lambda|$ the resolvent kernel coincides with the analytic continuation of the sum of the Neumann series.

The same result is obtained if we begin with any other zeroth approximation because the difference between the approximations disappeiars in the limiting process.

The iteration method can be applied to practical calculations of numerical solutions if we take an appropriate finite number of iterations; it can also be used for improving an approximate solution obtained with the help of some other method. However, these applications are only possible within the domain of convergence of series (45). In practice we judge upon the convergence by comparing successive approximations which must become closer to each other as the number of iterations increases. But outside the domain of convergence this technique is inapplicable.

6. Contraction Mapping Principle. The convergence of the iteration process for equation (16) with small $|\lambda|$ is a good illustration of the contraction mapping principle which has been widely used in recent years for testing the convergence of iterations for a large variety of problems. It turns out that the techniques used for testing convergence are essentially the same in many cases, which leads to the formulation of a general principle embracing the essential features of such techniques.

A natural range of application of this principle is a metric space. We remind the reader (Sec. VI.3.7) that a metric space is a set (R) of objects (or elements or points) for which the notion of a distance $\rho(\alpha, \beta)$ between two points $\alpha, \beta \in (R)$ is introduced. The distance must satisfy the three axioms given in Sec. VI.3.7. Euclidean spaces of arbitrary dimension, normed spaces and, in particular, normed functional spaces (Sec. VI.1.3) and any sets of

elements in them are examples of metric spaces. The notion of a metric space is not connected with linearity, and therefore, for instance, Riemannian spaces (Sec. V.2.5) are also metric, etc.

The notion of convergence can be introduced in a natural way for a metric space ($\alpha_n \xrightarrow[n \to \infty]{} \alpha$ means that $\rho(\alpha_n, \alpha) \xrightarrow[n \to \infty]{} 0$) as well as many other notions related to convergence which we are not going to discuss here. Any set (R_1) of points of a metric space (R) forms a metric space if we take the same distance function in (R_1) as in (R) (why?). A metric space (R_1) of this type is termed the **subspace of the space** (R).

Many (but by far not all) metric spaces possess an important property of completeness mentioned in Sec. VI.1.3. (It should be noted that this property has nothing in common with the completeness of the system of coordinate functions which was considered in Sec. VI.4.1.) The completeness of a metric space has a number of definitions which, as it can be proved, are equivalent. For instance:

1. Any sequence $\alpha_1, \alpha_2, \alpha_3 \ldots$ of points of (R) for which $\rho(\alpha_n, \alpha_m) \xrightarrow[m, n \to \infty]{} 0$ converges to a point belonging to (R).

2. The same but on condition that $\sum_{k=1}^{\infty} \rho(\alpha_k, \alpha_{k+1}) < \infty$.

3. A sequence of points of (R) cannot be convergent to a point not belonging to (R) and belonging to a space containing (R) as a subspace.

We shall enumerate some complete spaces (the proof of their completeness can be found in courses of functional analysis). These are all finite-dimensional Euclidean spaces, the space $C[a, b]$ and, in general, any space $C(K)$ of all real or complex (scalar or vector) continuous functions on a closed set (K) of a finite-dimensional Euclidean space if the distance between two functions is understood as their uniform deviation. (The set is closed if it contains the limits of all convergent sequences of its points.) By the way, for the spaces C the second definition of completeness coincides with the condition of Weierstrass' test for uniform convergence (e.g. see [107], Sec. XVII.8), which we leave to the reader to verify. Other examples of a complete space are the space $L_2[a, b]$ and, generally, the Hilbert space $L_2(M)$ (which is defined similarly) for an arbitrary set (M) with a nonnegative measure and, in particular, for any manifold (M) in a finite-dimensional Euclidean space with ordinary Lebesgue's measure. There are many other complete spaces which we are not going to consider here. We only mention that a closed (in the above sense) subspace of a complete metric space is complete (prove it!). At the same time all nonclosed subspaces are not complete; for instance, the set of all continuous functions

in $L_2[a, b]$ and the set of all continuously differentiable functions in $C[a, b]$ are not complete metric spaces (why?).

Now let us take a mapping B of a complete metric space (R) into itself. This means that to each point $\alpha \in (R)$ there corresponds a point $B\alpha \in (R)$. B is called a **contraction mapping** if there exists a positive number $k < 1$ such that

$$\rho(B\alpha, B\beta) \leqslant k\rho(\alpha, \beta) \tag{46}$$

for any $\alpha, \beta \in (R)$. (Suppose $k = \frac{1}{2}$, then condition (46) means that under the mapping B all distances in (R) become at least twice as short.) A contraction mapping is always continuous, i.e. $\alpha_n \xrightarrow[n\to\infty]{} \alpha$ implies $B\alpha_n \xrightarrow[n\to\infty]{} B\alpha$ (why?). The converse statement is not of course true; for instance, the mapping $y = x^2$, i.e. $x \longrightarrow x^2$, of the whole real axis is continuous but not contracting (though for the interval $-\frac{1}{3} \leqslant x \leqslant \frac{1}{3}$ of the real axis it is contracting).

Let us consider the "equation"

$$\alpha = B\alpha \tag{47}$$

In other words, let us look for a point α which goes into itself under the mapping B. Such a point is known as a **fixed point** of the mapping. It is natural to seek this point by iteration. We choose a zeroth approximation α_0 and construct the subsequent approximations according to the formulas $\alpha_1 = B\alpha_0$, $\alpha_2 = B\alpha_1$ and, generally,

$$\alpha_{n+1} = B\alpha_n, \quad n = 0, 1, 2, \ldots \tag{48}$$

The **contraction mapping principle** states that *if the space* (R) *is complete and* B *is a contraction mapping of* (R) *into itself, the sequence of approximations* α_n *constructed by formula* (48) *converges to a fixed point of the mapping* B *and this fixed point is unique.* (Observe that this principle gives only a sufficient condition for convergence of iterations.)

The principle is easily proved. Applying inequality (46) we obtain, in succession,

$$\rho(\alpha_1, \alpha_2) = \rho(B\alpha_0, B\alpha_1) \leqslant k\rho(\alpha_0, \alpha_1),$$
$$\rho(\alpha_2, \alpha_3) = \rho(B\alpha_1, B\alpha_2) \leqslant k\rho(\alpha_1, \alpha_2) \leqslant k^2\rho(\alpha_0, \alpha_1), \text{ etc.}$$

The series composed of the right-hand members of these relations being convergent, the sequence $\alpha_0, \alpha_1, \alpha_2, \ldots$ converges, according to the second definition of completeness, to a point $\bar{\alpha}$. Passing

in equation (48) to the limit for $n \to \infty$ we see that $\bar{\alpha}$ satisfies equation (47), that is it is a fixed point. The assumption that there exists another fixed point $\bar{\bar{\alpha}} \neq \bar{\alpha}$ results in the inequality

$$\rho(\bar{\alpha}, \bar{\bar{\alpha}}) = \rho(B\bar{\alpha}, B\bar{\bar{\alpha}}) \leqslant k\rho(\bar{\alpha}, \bar{\bar{\alpha}})$$

and thus leads to a contradiction. All the assertions have thus been proved. By the way, this example shows that a general consideration may prove to be simpler than the investigation of a special case because in the former the essential features of the problem are taken in a pure form whereas in the latter they may be obscured by the particulars.

To apply this principle to an equation of a special type, one should write the equation in form (47), specify in what space the mapping B operates and verify that the space is complete and that the mapping is contracting. Let us take equation (16) and consider it in the complete space $L_2[a, b]$. Then we put

$$Bu(x) = \lambda \int_a^b K(x, \xi)\, u(\xi)\, d\xi + f(x)$$

By Sec. 1.3, if the kernel satisfies condition (1.7) and $f \in L_2[a, b]$ then $Bu \in L_2[a, b]$, and hence B is a mapping of the space $L_2[a, b]$ into itself. Furthermore, we have

$$[\rho(Bu, Bv)]^2 = \| Bu - Bv \|^2 = \int_a^b | Bu(x) - Bv(x) |^2 dx =$$

$$= |\lambda|^2 \int_a^b \left| \int_a^b K(x, \xi) [u(\xi) - v(\xi)]\, d\xi \right|^2 dx$$

Applying inequality (1.11) we see that the right-hand member does not exceed

$$|\lambda|^2 \int_a^b \left(\int_a^b |K(x, \xi)|^2 d\xi \right) \left(\int_a^b |u(\xi) - v(\xi)|^2 d\xi \right) dx =$$

$$= |\lambda|^2 \int_a^b \int_a^b |K(x, \xi)|^2 dx\, d\xi\, [\rho(u, v)]^2$$

Thus, if

$$|\lambda| < \left(\int_a^b \int_a^b |K(x, \xi)|^2 dx\, d\xi \right)^{-\frac{1}{2}} \tag{49}$$

the conditions of the contraction mapping principle are fulfilled and the iterations constructed in Sec. 5 are convergent to the unique solution of equation (16). Incidentally, we obtain an esti-

mate for the least of the moduli of the characteristic values of equation (47): this modulus is not less than the right-hand term in (49) (why?).

It should be stressed that inequality (49) is only a sufficient condition for the solvability of equation (16) and the convergence of iterations. The convergence usually takes place in a larger circle of the λ-plain whose radius is equal to the least of the moduli of the characteristic values of the kernel K. Taking different norms in functional spaces and considering the iterated equation we can derive various conditions for convergence which often prove more convenient than (49).

7. Perturbed Kernels. If a characteristic value and the eigenfunctions associated with it are known for a given kernel, the small-parameter method makes it possible to determine the corresponding characteristic values and the eigenfunctions for "perturbed kernels" which are close to that kernel.

Consider the equation

$$u(x) = \lambda \int_a^b K(x, \xi, \mu)\, u(\xi)\, d\xi \tag{50}$$

where μ is a small real parameter. Suppose that a characteristic value λ_0 of the kernel $K_0(x, \xi) = K(x, \xi, 0)$ is known and that to this value there corresponds only one eigenfunction $u_0(x)$ ($u_0(x) \not\equiv 0$). To find the corresponding characteristic value λ and the eigenfunction $u(x)$ for $\mu \neq 0$ we take the expansions

$$K(x, \xi, \mu) = K_0(x, \xi) + \mu K_1(x, \xi) + \mu^2 K_2(x, \xi) + \ldots,$$
$$\lambda = \lambda_0 + \mu\lambda_1 + \mu^2\lambda_2 + \ldots, \quad u(x) = u_0(x) + \mu u_1(x) + \mu^2 u_2(x) + \ldots \tag{51}$$

Substituting (51) into (50) and equating coefficients in like powers of μ we obtain the equality

$$u_0(x) = \lambda_0 \int_a^b K_0(x, \xi)\, u_0(\xi)\, d\xi \tag{52}$$

(which is a direct consequence of the definition of λ_0 and $u_0(x)$) and the relations

$$u_1(x) = \lambda_0 \int_a^b K_0(x, \xi)\, u_1(\xi)\, d\xi +$$
$$+ \lambda_0 \int_a^b K_1(x, \xi)\, u_0(\xi)\, d\xi + \lambda_1 \int_a^b K_0(x, \xi)\, u_0(\xi)\, d\xi \tag{53}$$

$$u_2(x) = \lambda_0 \int_a^b K_0(x, \xi)\, u_2(\xi)\, d\xi + \lambda_0 \int K_1 u_1\, d\xi +$$
$$+ \lambda_0 \int K_2 u_0\, d\xi + \lambda_1 \int K_0 u_1\, d\xi + \lambda_1 \int K_1 u_0\, d\xi + \lambda_2 \int K_0 u_0\, d\xi \tag{54}$$

etc. By Fredholm's third theorem, we conclude from equality (53) that the sum of the terms containing u_0 must be orthogonal to the eigenfunction $v_0(x)$ of the equation

$$\left(\lambda_0 \int_a^b K_1(x, \xi)\, u_0(\xi)\, d\xi + \lambda_1 \int_a^b K_0(x, \xi)\, u_0(\xi)\, d\xi,\ v_0(x)\right) = 0$$

adjoint to (52). Whence, using equality (52) we deduce

$$\lambda_1 = \left(-\lambda_0^2 \int_a^b \int_a^b K_1(x, \xi)\, u_0(\xi)\, v_0^*(x)\, d\xi\, dx\right) \times \times \left(\int_a^b u_0(x)\, v_0^*(x)\, dx\right)^{-1}$$

The coefficient λ_1 satisfying the orthogonality condition being found, we thus obtain the general solution of equation (53) in the form $u_1(x) = \varphi_1(x) + C_1 u_0(x)$ where C_1 is an arbitrary constant. This arbitrariness is inessential and is due to the possibility of multiplying $u(x)$ by any expansion $1 + C_1\mu + C_2\mu^2 + \ldots$. Therefore, we choose C_1 according to some normalization condition for the solution or simply put it equal to zero. Now we take equation (54) from which, by the orthogonality of the nonhomogeneity term to the function v_0, we find λ_2 and then determine $u_2(x)$, and so on.

If a characteristic value λ_0 of the kernel K_0 is multiple, then, for $\mu \neq 0$, it may generate several characteristic values of the kernel K. In this case expansion (51) should contain fractional powers of the parameter μ and the calculations become more complicated; here we shall not treat this case.

Multidimensional equations with perturbed kernels are investigated in a similar way. Of interest also are perturbed equations of the form

$$u(M) = \lambda \int_{(G_\mu)} K(M, N, \mu)\, u(N)\, d_N G$$

in which not only the kernel is dependent on a parameter but the domain of integration as well. In this case it may be convenient to introduce a change of variables $M' = \varphi(M, \mu)$ defining a one-to-one mapping of the domain (G_μ) onto a domain (G') independent of μ; the problem is then reduced to the above problem. We can avoid such a change if it is permissible to differentiate the integral with respect to the parameter specifying the shape of the domain of integration. Then, for example, if the domain (G_μ) is determined by the inequality $\varphi_0(M) + \mu\varphi_1(M) + \ldots \leqslant 0$ the right-

hand side of relation (53) (taken for the multidimensional case) will contain an additional term of the form

$$-\lambda_0 \int\limits_{(S_0)} K_0(M, N) \frac{\varphi_1(N)}{|\operatorname{grad} \varphi_0(N)|} u_0(N)\, d_N S_0$$

where (S_0) is the boundary of the domain $(G_0) = (G_\mu)|_{\mu=0}$ (e.g. see [107], Sec. XVI.18).

8. **Properties of Solutions.** Though, generally, we are interested in solutions of equation (16) belonging to the space $L_2[a, b]$ whose elements may be discontinuous we often encounter cases when the solution is nevertheless continuous and even has a continuous derivative. For example, let a and b be finite. Suppose that the kernel $K(x, \xi)$ and the function $f(x)$ are continuous in their arguments. Then it is easy to prove that the solution $u(x)$ is also continuous. Indeed, applying the inequalities (1.11) and $|a+b|^2 \leqslant \leqslant 2|a|^2+2|b|^2$ we obtain

$$|u(x+\Delta x)-u(x)|^2 = |\lambda \int\limits_a^b [K(x+\Delta x, \xi)- -K(x, \xi)]\, u(\xi)\, d\xi + [f(x+\Delta x)-f(x)]|^2 \leqslant$$

$$\leqslant 2|\lambda|^2 \int\limits_a^b |K(x+\Delta x, \xi)-K(x, \xi)|^2 d\xi \int\limits_a^b |u(\xi)|^2 d\xi + +2|f(x+\Delta x)-f(x)|^2$$

whence it follows that the function $u(x)$ is continuous. (If, for a continuous kernel, the function $f(x)$ is discontinuous, it can be similarly verified that the difference $u(x)-f(x)$ is continuous, that is the solution has the same discontinuities as $f(x)$.)

Now suppose that the kernel $K(x, \xi)$ has points or lines of discontinuity and the function $f(x)$ is continuous. Then the solution $u(x)$ may have discontinuities only at those points $x=x_0$ for which the straight line $x=x_0$ contains some segments belonging to the lines of discontinuity of the kernel. See Fig. 114 where the lines of discontinuity of a kernel are shown in heavy lines. Since the integration in (16) is carried out along vertical segments, the integral may gain a finite increment when we pass from $x=x_1<x_0$ to a close value $x=x_2>x_0$, which means that the solution has a discontinuity.

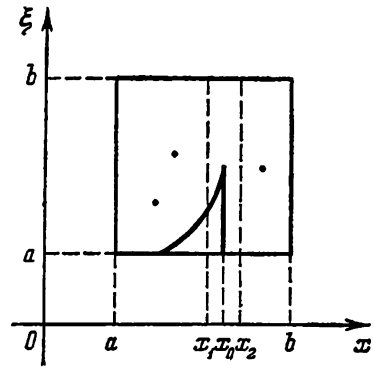


Fig. 114

In particular, if the lines of discontinuity of the kernel do not contain segments parallel to the ξ-axis and the function $f(x)$ is continuous the solution $u(x)$ is also continuous.

Let us consider a typical case when the kernel and its derivatives are discontinuous for $x=\xi$ and continuous for $x\neq\xi$. Then, as a rule, the kernel has different analytical expressions for $\xi<x$ and for $\xi>x$, and therefore it is convenient to write it in the form

$$K(x, \xi)=\begin{cases} K_1(x, \xi) & \text{for} \quad \xi<x \\ K_2(x, \xi) & \text{for} \quad \xi>x \end{cases}$$

where K_1 and K_2 are smooth functions. Equation (16) can be rewritten as

$$u(x)=\lambda\int_a^x K_1(x, \xi)\,u(\xi)\,d\xi+\lambda\int_x^b K_2(x, \xi)\,u(\xi)\,d\xi+f(x)$$

This again indicates that if the functions

$$K_1(x, \xi)\ (a\leqslant\xi\leqslant x\leqslant b),\ K_2(x, \xi)\ (a\leqslant x\leqslant\xi\leqslant b),\ f(x)\ (a\leqslant x\leqslant b)$$

are continuous the solution $u(x)$ is continuous as well. If, in addition, the functions K_1, K_2 and f possess continuous derivatives with respect to x then, calculating the derivative of the integral with respect to the parameter (e.g. see [107], Sec. XIV.20), we obtain

$$u'(x)=\lambda[K_1(x, x)-K_2(x, x)]\,u(x)+$$
$$+\lambda\int_a^x K'_{1x}(x, \xi)\,u(\xi)\,d\xi+\lambda\int_x^b K'_{2x}(x, \xi)\,u(\xi)\,d\xi+f'(x)$$

which implies that the solution is continuously differentiable in x. We can similarly indicate the conditions on K_1, K_2 and f under which u'' is continuous, etc. Thus, the degree of smoothness of the solution is specified by the properties of the kernel and the function $f(x)$.

The above results are also extended to n-dimensional equations of form (1.9). In particular, if the kernel $K(M, N)$ has a discontinuity only for $M=N$, the integral in (1.9) depends on the point M continuously, and therefore if the nonhomogeneity term $f(M)$ is continuous the solution $u(M)$ is also continuous. Integrals of this type are encountered in mathematical physics and are known as **integrals of the potential type** because the Newtonian and the logarithmic potentials (Sec. 1.2.5) are their particular cases. Mathematical physics also establishes conditions guaranteeing differentiability of such integrals but here we shall not consider this question.

9. Volterra's Integral Equations of the Second Kind. Let us consider the Volterra linear integral equation of the second kind (Sec. 1.2)

$$u(x)=\lambda\int_a^x K(x,\xi)\,u(\xi)\,d\xi+f(x) \tag{55}$$

where a is finite. It can be proved that *this equation has no characteristic values*, which means that for every λ equation (55) has exactly one solution (i.e. in this case Fredholm's first alternative takes place; see Sec. 2).

Indeed, suppose the contrary; then, by Fredholm's first theorem, the corresponding homogeneous equation

$$u(x)=\lambda\int_a^x K(x,\xi)\,u(\xi)\,d\xi \tag{56}$$

has a nontrivial solution. If x varies from a to some $a+h>a$ we can put, conditionally,

$$K(x,\xi)\equiv 0 \quad \text{for} \quad \xi>x \tag{57}$$

and derive from (1.13) the inequality

$$\int_a^{a+h}|u(x)|^2\,dx\leqslant|\lambda|^2\int_a^{a+h}\int_a^{a+h}|K(x,\xi)|^2\,dx\,d\xi\int_a^{a+h}|u(x)|^2\,dx$$

If h is so small that the product of the first two factors on the right-hand side is less than unity, it follows that $u(x)\equiv 0$ for $a\leqslant x\leqslant a+h$ (why?). Furthermore, if the solution $u(x)$ becomes different from zero for $x\geqslant a_1>a$, then, replacing in (56) the lower limit of integration by a_1, and reasoning in the same way, we arrive at a contradiction. Thus, the above assertion has been proved.

According to the construction of iterated kernels in the iteration method we obtain from (57) the expression

$$K_2(x,\xi)=\int_a^{\xi}K(x,\eta)\,K(\eta,\xi)\,d\eta+\int_{\xi}^{x}K(x,\eta)\,K(\eta,\xi)\,d\eta+$$

$$+\int_x^b K(x,\eta)\,K(\eta,\xi)\,d\eta=\int_{\xi}^{x}K(x,\eta)\,K(\eta,\xi)\,d\eta$$

whence it follows that the kernel K_2 also satisfies condition (57). Considering K_3, etc. we see that all the iterated kernels satisfy condition (57), and therefore the integral in formula (43) can be taken from ξ to x. Consequently, formula (44) expressing the solu-

tion can be written in the form

$$u(x)=f(x)+\lambda\int_a^x \Gamma(x, \xi, \lambda) f(\xi)\, d\xi \tag{58}$$

Besides, the resolvent kernel of Volterra's equation regarded as a function of λ has no singular points, and therefore series (45) and, together with it, the iteration process are convergent for all finite values of λ.

We shall denote the independent variable by the letter t and interpret it as time. Then we see that for Volterra's operator

$$Au(t)=\int_{t_0}^{t} K(t, \tau)\, u(\tau)\, d\tau \qquad (t \geqslant t_0)$$

(in contrast to Fredholm's general operator) the value of the function Au at any moment t is determined solely by the values of the function u for $\tau \leqslant t$. The operators (including nonlinear ones) possessing this property are called **operators of the Volterra type.** They are widely used for investigating processes with aftereffect (why?). It is natural that solution (58) is also completely specified at each moment only by the values of the external action f at the preceding moments of time.

It is characteristic of Volterra's equation of the second kind (55) that *it is possible to extend its solution.* For instance, we can first construct the solution on an interval $a \leqslant x \leqslant a_1$, then rewrite the equation for $x \geqslant a_1$ in the form

$$u(x)=\lambda\int_{a_1}^{x} K(x, \xi)\, u(\xi)\, d\xi+\left[\lambda\int_a^{a_1} K(x, \xi)\, u(\xi)\, d\xi+f(x)\right]$$

and continue to solve it since the sum in the square brackets has already been found, etc. For Fredholm's general equation (16) the solution cannot be constructed solely on a part of the interval $[a, b]$ because, according to the meaning of the equation, all the values of the sought-for function are interconnected.

There are different ways of generalizing the notion of Volterra's equation to functions of several variables. For example, we can consider the equation

$$u(x_1, x_2, \ldots, x_n)=\lambda\int_{a_1}^{x_1} d\xi_1 \underbrace{\int\ldots\int}_{(H)} K(x_1, x_2, \ldots,$$

$$x_n, \xi_1, \xi_2, \ldots, \xi_n) \times u(\xi_1, \xi_2, \ldots, \xi_n)\, d\xi_2 \ldots d\xi_n$$

where (H) is a domain in the $x_2, \ldots, x_n$-space. Such equations have no characteristic values either.

10. Equations with Polar Singularities. Fredholm's theorems can be extended to equation (16) with a kernel satisfying condition (1.8) and finite a and b. We shall briefly show how this can be done. First of all let us prove that if $u(x) \in L_2[a, b]$, assertion (1.12) remains true although the former proof based on (1.13) is no longer applicable (why?). Indeed, we can write (check it up!)

$$\int_a^b \left| \int_a^b K(x, \xi)\, u(\xi)\, d\xi \right|^2 dx \leqslant C \int_a^b \left| \int_a^b \frac{1}{|x-\xi|^{\frac{\alpha}{2}}} \frac{u(\xi)}{|x-\xi|^{\frac{\alpha}{2}}} d\xi \right|^2 dx \leqslant$$

$$\leqslant C \int_a^b \left(\int_a^b \frac{1}{|x-\xi|^{\alpha}} d\xi \int_a^b \frac{|u^2(\xi)|}{|x-\xi|^{\alpha}} d\xi \right) dx \leqslant C_1 \int_a^b dx \int_a^b \frac{|u^2(\xi)|}{|x-\xi|^{\alpha}} d\xi =$$

$$= C_1 \int_a^b d\xi\, |u^2(\xi)| \int_a^b \frac{1}{|x-\xi|^{\alpha}} dx \leqslant C_2 \int_a^b |u^2(\xi)|\, d\xi$$

which implies (1.12). Thus, for a kernel with a polar singularity the operator A introduced in Sec. 1.3 is bounded.

The main principle of the theory of kernels with polar singularity is based on the fact that for the nth iterated kernel the exponent α characterizing the singularity decreases as the number n increases, and so, beginning with some n, the singularity disappears. This is what makes it possible to pass from the original equation to the equation for which the validity of Fredholm's theorems has already been proved.

Let us verify the above assertion concerning singularities. We suppose that a kernel $K(x, \xi)$ satisfies inequality (1.8) and that another kernel $L(x, \xi)$ satisfies a similar inequality with exponent β. Then the kernel resulting from the convolution of K and L satisfies the inequality

$$\left| \int_a^b K(x, \eta)\, L(\eta, \xi)\, d\eta \right| \leqslant C \int_a^b |x-\eta|^{-\alpha} |\eta-\xi|^{-\beta} d\eta \leqslant$$

$$\leqslant C |x-\xi|^{1-\alpha-\beta} \left| \int_{\frac{(a-\xi)}{(x-\xi)}}^{\frac{(b-\xi)}{(x-\xi)}} |1-s|^{-\alpha} |s|^{-\beta} ds \right| \quad (\eta - \xi = s(x-\xi)) \qquad (59)$$

Since for $s=0$ and $s=1$ the last integral is convergent, and for $|s| \to \infty$ the integrand is equivalent to $|s|^{-(\alpha+\beta)}$, the integral does not exceed $C_1 + C_2 |x-\xi|^{\alpha+\beta-1}$ in its modulus (why?). Substituting the last quantity into (59) we obtain

$$\left| \int_a^b K(x, \eta)\, L(\eta, \xi)\, d\eta \right| \leqslant \begin{cases} \dfrac{\text{const}}{|x-\xi|^{\alpha+\beta-1}} & \text{if } \alpha+\beta-1 > 0, \\ \text{const} & \text{if } \alpha+\beta-1 < 0 \end{cases} \qquad (60)$$

(The case $\alpha+\beta=1$ does not involve an additional consideration because in inequality (1.8) we can always give α a small increment.)

From the above proof it follows that the kernel K_2 has a singularity of order $\alpha-(1-\alpha)$ if $\alpha-(1-\alpha)>0$, K_3 has a singularity of order $\alpha-2(1-\alpha)$ if $\alpha-2(1-\alpha)>0$, etc. Finally, when we arrive at m such that $\alpha-m(1-\alpha)<0$ we obtain a kernel K_{m-1} which is bounded.

It should be noted that a similar result is valid for n-dimensional integrals, but in this case it is necessary to replace $\alpha+\beta-1$ by $\alpha+\beta-n$ in (60).

Now let us suppose, for simplicity, that $\alpha-(1-\alpha)<\frac{1}{2}$; then the kernel K_2 satisfies inequality (1.7). Equation (16) can be written in the shortened form

$$(I-\lambda A)\,\mathbf{u}=\mathbf{f} \tag{61}$$

where I is the *unit* (*identity*) operator (making no changes in all the functions). We apply the operator $I+\lambda A$ to both sides and obtain the equation

$$(I-\lambda^2A^2)\,\mathbf{u}=(I+\lambda A)\,\mathbf{f},\ \text{i. e. }\ \mathbf{u}=\lambda^2A^2\mathbf{u}+(I+\lambda A)\,\mathbf{f} \tag{62}$$

where the function $(I+\lambda A)\,\mathbf{f}$ is known and A^2 is the integral operator with the kernel K_2 for which Fredholm's theorems hold. Incidentally, it is seen from the above proof that every solution of equation (61) satisfies equation (62) as well. In particular, putting $f\equiv 0$ we see that if $\mathbf{u}$ is an eigenfunction of equation (61) corresponding to a characteristic value λ it is also an eigenfunction of equation (62) corresponding to the eigenvalue λ^2. It readily follows that since Fredholm's fourth theorem (Sec. 2) holds for equation (62) it holds for equation (61) as well. It also follows that for every characteristic value of equation (61) the space of the corresponding eigenfunctions is of a finite dimension (why?).

Henceforth we shall consider λ to be fixed and suppose that $-\lambda$ is not a characteristic value of equation (61). (In comprehensive courses on integral equations it is proved by means of iterations of higher orders that this condition is inessential for the final results to be valid.) Then it is easily proved that every solution of equation (62) satisfies equation (61) as well, that is equations (61) and (62) are equivalent. This is readily seen if equation (62) is rewritten in the form

$$[(I-\lambda A)\,\mathbf{u}-\mathbf{f}]=(-\lambda)\,A\,[(I-\lambda A)\,\mathbf{u}-\mathbf{f}]$$

(why?).

The homogeneous equation adjoint to equation (61) with a concrete value of λ involves the complex conjugate value λ^*

of the parameter and has the form

$$(I-\lambda^* A^*)\,v=0 \tag{63}$$

By analogy with the above it can be shown that (63) is equivalent to the equation

$$v=\lambda^{*2} A^{*2} v \tag{64}$$

and it can be verified directly that equation (64) is adjoint to (62) because the iterated adjoint kernel is adjoint to the iterated original kernel (why?). Applying Fredholm's third theorem to equation (62) we find that for its solvability (and, consequently, for the solvability of equation (61)) it is necessary and sufficient that the orthogonality condition

$$((I+\lambda A)\,f,\ v)=0 \tag{65}$$

be fulfilled for any solution of equation (64), i.e. of (63). However, it can readily be shown (do it!) that $(Af, v)=(f, A^*v)$, and therefore condition (65) can be rewritten in the form

$$0=(f, v)+(\lambda Af, v)=(f, v)+(f, \lambda^* A^* v)=2\,(f, v)$$

We thus arrive at Fredholm's third theorem for equation (61). As was indicated in Sec. 1, Fredholm's first theorem is implied by the third and second theorems.

Thus, the Fredholm four theorems hold for equations with polar singularity as well.

11. Equations with Completely Continuous Operators. A detailed analysis of Fredholm's theorems given in courses on integral equations and functional analysis shows that they hold in some more general cases. Here we shall present these general facts without proof.

Let us consider a Hilbert space (H). It may be a functional space which we discussed in Sec. 1.3 or, simply, an infinite-dimensional Euclidean space with elements whose nature is inessential. Consider a linear operator A which specifies a mapping of (H) into itself. Such an operator is said to be **finite-dimensional** if it maps (H) into its finite-dimensional subspace. This definition generalizes the notion of an integral operator with a degenerate kernel because the operator

$$Au(x)=\int_a^b \sum_{j=1}^{p} \Phi_j(x)\,\Psi_j(\xi)\,u(\xi)\,d\xi$$

maps the space $L_2[a, b]$ onto the subspace formed of all the linear combinations of the functions $\Phi_1(x)$, $\Phi_2(x)$, ..., $\Phi_p(x)$ (why?). Finite-dimensional operators are always bounded (see the definitions of a bounded operator and of the norm of a bounded operator in Sec. 1.3).

A linear operator A *is called* **completely continuous (compact)** *if it can be approximated, to an arbitrary accuracy, with a finite-dimensional operator*, that is, if it can be represented as a sum of a finite-dimensional operator and an operator with arbitrarily small norm. For instance, as it follows from the beginning of Sec. 2, integral operators of the Fredholm type are completely continuous. It can be proved that integral operators with polar singularity are also completely continuous.

The following equivalent definition of a completely continuous operator is also frequently used: *an operator* A *is said to be completely continuous if it maps every bounded set onto a compact set of* (H). (The definition of a compact set was given in Sec. VI.2.9 where it was also shown that a bounded set of an infinite-dimensional space may not be compact. Moreover, it can be proved that only in finite-dimensional spaces every bounded set is compact.) Hence, it follows that every completely continuous operator is bounded, but in the general case the converse is not true. For instance, a unit operator in any infinite-dimensional space is bounded but not completely continuous.

For every linear operator A it is possible to construct its adjoint linear operator A^* which also maps (H) into (H) and is connected with A by the relationship

$$(\mathbf{Ax},\ \mathbf{y}) = (\mathbf{x},\ \mathbf{A^*y}) \text{ for all } \mathbf{x},\ \mathbf{y} \in H \qquad (66)$$

(see an analogous definition in Sec. IV.1.3). In the case of integral operators, A^* is nothing but the operator corresponding to the adjoint (associated) kernel (see formula (18)). If A is completely continuous, A^* is completely continuous as well.

F. Riesz (1880-1956), a Hungarian mathematician, extended Fredholm's theory to operator equations

$$\mathbf{u} = \lambda \mathbf{A}\mathbf{u} + \mathbf{f} \qquad (67)$$

where $\mathbf{f}$ is a given element of Hilbert's space (H), $\mathbf{u}$ is the sought-for element and A is a completely continuous linear operator specifying a mapping of (H) into itself. It turns out that for such equations Fredholm's four theorems hold in exactly the same form as they were formulated in Sec. 1 (here, of course, instead of eigenfunctions we speak about generalized eigenvectors). In this case the adjoint equation to (67) is naturally written as $\mathbf{v} = \mu \mathbf{A}^* \mathbf{v}$.

Thus, it is possible to apply Fredholm's theorems if the corresponding operator is completely continuous. The verification of this assertion usually involves some sophisticated techniques studied in courses of functional analysis.

12. Equations with Positive Kernels. In many applications the kernel of equation (16) is nonnegative:

$$K(x,\ \xi) \geqslant 0 \qquad (a \leqslant x,\ \xi \leqslant b) \qquad (68)$$

Equations with such kernels in one-dimensional and multidimensional cases as well as nonlinear analogues of these equations possess a number of interesting properties discussed in detail in [67] and [139].

It turns out that all the assertions given in Sec. IV.4.8 for matrices with positive elements hold for equation (16) with a continuous positive kernel (for which inequality (68) is strict: $K(x, \xi) > 0$). But in this case instead of a vector with nonnegative projections we speak about a function assuming nonnegative values, and instead of the eigenvalue with the greatest modulus we speak about the characteristic value with the least modulus. These assertions also hold for multidimensional equations, discontinuous kernels and nonstrict inequality (68). In the latter case, however, the kernel must satisfy some additional conditions which can be found in the books mentioned above. Among the equations dealt with in most applications only Volterra's equations do not satisfy these conditions because they have no characteristic values (Sec. 9). The basic facts of the theory of oscillatory matrices discussed in Sec. IV.4.8 are also extended to integral equations.

The theory of the so-called operator equations in spaces with a cone is an abstract analogue of the theory of equations with positive kernels. The notion of a cone generalizes the notion of a set of nonnegative functions or of a set of vectors with nonnegative projections. More precisely, a **cone** in a Hilbert or a Banach space (H) is a closed set (K) of points (elements) possessing the following properties:

1. If $\mathbf{x} \in (K)$, $\mathbf{y} \in K$, $\alpha \geqslant 0$, $\beta \geqslant 0$ then $\alpha\mathbf{x} + \beta\mathbf{y} \in (K)$.
2. If $\mathbf{x} \in (K)$, $\mathbf{x} \neq 0$ then $-\mathbf{x} \overline{\in} (K)$.

(What do these properties mean for two- and three-dimensional Euclidean spaces?) For instance, in any functional space the set of all nonnegative functions is a cone (why?). A completely continuous linear operator (Sec. 11) which maps a cone (K) into itself serves as an analogue of an integral operator with a nonnegative kernel. For equation (67) in which the operator A possesses this property it is possible to prove, under certain assumptions, all the assertions stated in Sec. IV.4.8 for the finite-dimensional case.

The theory of spaces with a cone was founded by the Soviet mathematician M.G. Krein (born 1907).

§ 3. EQUATIONS WITH SYMMETRIC KERNELS

1. Finite-Dimensional Analogue. Let us again take equation (2.16) and suppose that $K(x, \xi)$ is a real symmetric kernel, that is one satisfying the condition

$$K(x, \xi) \equiv K(\xi, x) \tag{1}$$

Symmetric matrices possess a number of characteristic properties (e.g. see [107], Sec. XI.10 and Sec. IV.1.6 of this book). It turns out that these properties can also be proved for integral operators with symmetric kernels. When establishing the corresponding properties we shall simplify our consideration by assuming that it is possible to replace, approximately, integral equation (2.16) by a system of linear algebraic equations of type (2.34). For more rigorous proofs we refer the reader to comprehensive courses on integral equations.

Since for a symmetric kernel K the matrix of the coefficients of system (2.34) is symmetric, we conclude from the properties of such matrices that *all characteristic values of an equation with a real symmetric kernel are real.* Therefore, unless otherwise stated, we shall limit ourselves to considering real functions.

Denote the components of two eigenvectors of equation (2.34) corresponding to two different eigenvalues by $u_k^{(1)}$ and $u_k^{(2)}$. As is known from the theory of symmetric matrices, for these components the orthogonality relation

$$\sum_k u_k^{(1)} u_k^{(2)} = 0$$

is fulfilled. Multiplying both members by h and taking sufficiently small values of h we conclude that *the eigenfunctions of an equation with symmetric kernel corresponding to different characteristic values are mutually orthogonal.*

Furthermore, *every nonzero symmetric kernel* (i.e. one which is not identically equal to zero) *has at least one characteristic value.* Indeed, for a finite-dimensional approximation this means that a symmetric matrix whose not all elements are equal to zero has at least one nonzero eigenvalue. (For nonsymmetric matrices this is not true in the general case, which accounts for the fact that Volterra's equations have no characteristic values.) Indeed, a symmetric matrix can always be reduced to a diagonal form, the diagonal elements being equal to the eigenvalues; if all the eigenvalues are equal to zero, then, after the reduction, we obtain a zero matrix. But this means that the matrix is equal to zero before the reduction (why?), which contradicts the original hypothesis.

2. Expanding Kernels with Respect to Eigenfunctions. Let us arrange the characteristic values of a given symmetric kernel $K(x, \xi)$ into a finite or infinite sequence

$$\lambda_1, \lambda_2, \ldots, \lambda_k, \ldots \tag{2}$$

in which every value is repeated as many times as the dimension of the space of the eigenfunctions associated with it. By Fred-

holm's fourth theorem we conclude that if sequence (2) is infinite then $\lambda_k \longrightarrow \infty$.

Now, let us choose a concrete eigenfunction for every characteristic value. If there are several linearly independent eigenfunctions associated with a given characteristic value we choose them in such a way that they are mutually orthogonal, which is always possible (why?). Besides, after multiplying these functions by the appropriate constant factors we can make the norm of every function equal unity. Thus we obtain a normalized sequence of eigenfunctions

$$\varphi_1(x),\ \varphi_2(x),\ \ldots,\ \varphi_k(x),\ \ldots \tag{3}$$

every two terms of which are orthogonal to each other. Indeed, as was shown in Sec. 1, the functions corresponding to different characteristic values are orthogonal to one another, and the functions corresponding to the same value are orthogonal by the construction.

Consider first the simplest case when the kernel has only one characteristic value λ_1 and suppose that there is only one linearly independent (i.e. nonzero) eigenfunction $\varphi_1(x)$ belonging to it. Then it is easily proved that

$$K(x,\ \xi) \equiv \frac{1}{\lambda_1}\varphi_1(x)\,\varphi_1(\xi) \tag{4}$$

Indeed, let us show that the auxiliary symmetric kernel

$$K_0(x,\ \xi) = K(x,\ \xi) - \frac{1}{\lambda_1}\varphi_1(x)\,\varphi_1(\xi) \tag{5}$$

has no characteristic values. By Sec. 1, this will imply that $K_0 \equiv 0$, which is what we set out to prove. To this end, suppose that $\varphi(x) \not\equiv 0$ is an eigenfunction of kernel (5) corresponding to a characteristic value λ:

$$\varphi(x) = \lambda \int_a^b K_0(x,\ \xi)\,\varphi(\xi)\,d\xi =$$

$$= \lambda \int_a^b K(x,\ \xi)\,\varphi(\xi)\,d\xi - \frac{\lambda}{\lambda_1}\varphi_1(x)\int_a^b \varphi(\xi)\,\varphi_1(\xi)\,d\xi \tag{6}$$

Multiplying both members by $\varphi_1(x)$, integrating from a to b, taking into account the symmetry of the kernel K which implies the equality

$$\int_a^b K(x,\ \xi)\,\varphi_1(x)\,dx = \int_a^b K(\xi,\ x)\,\varphi_1(x)\,dx = \frac{1}{\lambda_1}\varphi_1(\xi)$$

and using relation (6) we derive

$$\int_a^b \varphi(x)\,\varphi_1(x)\,dx = 0 \tag{7}$$

But, by virtue of (7), formula (6) then shows that $\varphi(x)$ is an eigenfunction of the kernel K and thus is proportional to $\varphi_1(x)$, which contradicts (7). Formula (4) has been proved.

Now, taking the general case when the kernel K has any number of eigenfunctions we can similarly prove that the system of characteristic values and eigenfunctions of kernel (5) is obtained if we delete the first term of sequences (2) and (3). (To prove this, it is necessary to verify that, for $j > 1$, every function $\varphi_j(x)$ is an eigenfunction of the kernel K_0 corresponding to the characteristic value λ_j, which we leave to the reader.) Then we can apply a similar transformation to kernel (5), etc. In particular, this procedure shows that *if a symmetric kernel has a finite number of characteristic values it is degenerate.*

If K is a nondegenerate kernel we arrive at the series

$$\sum_{j=1}^{\infty} \frac{1}{\lambda_j} \varphi_j(x)\,\varphi_j(\xi)$$

It can be shown that if condition (1.7) is fulfilled, this series is convergent in the mean but here we shall not dwell on the proof. Reasoning as in the case of kernel (5) we similarly conclude that the kernel

$$K(x, \xi) - \sum_{j=1}^{\infty} \frac{1}{\lambda_j} \varphi_j(x)\,\varphi_j(\xi)$$

possesses no characteristic values and thus arrive at the following formula for **the expansion of a symmetric kernel with respect to its eigenfunctions:**

$$K(x, \xi) = \sum_{j} \frac{1}{\lambda_j} \varphi_j(x)\,\varphi_j(\xi) \tag{8}$$

Here we do not write the limits for the summation index because they can be either finite in the case of a degenerate kernel or infinite in the case of a nondegenerate one.

3. Corollaries. Expansion (8) implies a number of useful formulas. For example, let us square both members of equality (8) and integrate the result with respect to x and with respect to ξ from a to b. In accordance with the orthogonality of the eigenfunctions, all the integrals of the products of different eigenfunctions vanish

and, consequently, by the normalization condition, we obtain the formula

$$\int_a^b \int_a^b [K(x, \xi)]^2 \, dx \, d\xi = \sum_j \frac{1}{\lambda_j^2} \tag{9}$$

Furthermore, if in (8) we put $\xi = x$ and integrate the result we obtain the formula

$$\int_a^b K(x, x) \, dx = \sum_j \frac{1}{\lambda_j} \tag{10}$$

The left-hand member of (10) is referred to as the **trace** of the continuous kernel K.

For the iterated kernel K_2 (see Sec. 2.5) we obtain formula

$$K_2(x, \xi) = \int_a^b K(x, \eta) K(\eta, \xi) \, d\eta =$$

$$= \int_a^b \sum_j \frac{1}{\lambda_j} \varphi_j(x) \varphi_j(\eta) \sum_k \frac{1}{\lambda_k} \varphi_k(\eta) \varphi_k(\xi) \, d\eta = \sum_j \frac{1}{\lambda_j^2} \varphi_j(x) \varphi_j(\xi) \tag{11}$$

and similarly derive the relation (check it up!)

$$K_n(x, \xi) = \sum_j \frac{1}{\lambda_j^n} \varphi_j(x) \varphi_j(\xi) \tag{12}$$

Since $\lambda_j \xrightarrow[j \to \infty]{} \infty$, the convergence of the series on the right-hand side of (12) is accelerated as n increases. By analogy with the preceding paragraph, we deduce from (12) the formulas

$$\int_a^b \int_a^b [K_n(x, \xi)]^2 \, dx \, d\xi = \sum_j \frac{1}{\lambda_j^{2n}}, \quad \int_a^b K_n(x, x) \, dx = \sum_j \frac{1}{\lambda_j^n} \tag{13}$$

Formulas (13) can be used for approximate evaluation of the first eigenvalues. Indeed, suppose that $|\lambda_1| < |\lambda_2| < \ldots$. Then, as n increases, in the right-hand sides of (13) the first summand becomes dominant. Discarding the other terms we obtain the approximate formula

$$|\lambda_1| \approx \left[\int_a^b \int_a^b [K_n(x, \xi)]^2 \, dx \, d\xi \right]^{-\frac{1}{2n}} \approx \left| \int_a^b K_n(x, x) \, dx \right|^{-\frac{1}{n}}$$

the sign of λ_1 being coincident with the sign of the second integral for odd values of n. It is also possible to use the ratios of

the traces:

$$\lambda_1 \approx \frac{\int_a^b K_n(x, x)\,dx}{\int_a^b K_{n+1}(x, x)\,dx}$$

To calculate λ_2 we can take advantage of the fact that in the expansion of the quantity

$$B_n = \left[\int_a^b K_n(x, x)\,dx\right]^2 - \int_a^b \int_a^b [K_n(x, \xi)]^2\,dx\,d\xi$$

the dominant term is $\frac{2}{\lambda_1^n \lambda_2^n}$ (why?) whence

$$|\lambda_2| \approx \frac{1}{|\lambda_1|} \sqrt[n]{\frac{2}{|B_n|}}$$

The **Kellog method** for approximate computation of λ_1 which we demonstrate below is also based on iteration. Let us take an arbitrary function $f(x)$ and put

$$f_1(x) = \int_a^b K(x, \xi) f(\xi)\,d\xi, \quad f_2(x) = \int_a^b K(x, \xi) f_1(\xi)\,d\xi$$

etc. This readily leads to the formulas (check it up!)

$$f_1(x) = \sum_j \frac{f_j}{\lambda_j} \varphi_j(x), \quad f_2(x) = \sum_j \frac{f_j}{\lambda_j^2} \varphi_j(x), \quad \ldots$$

where

$$f_j = \int_a^b f(x) \varphi_j(x)\,dx$$

If $f_1 \neq 0$, then for large n the principal role in $f_n(x)$ is played by the first summand. Therefore, we can use the approximate formulas

$$\lambda_1 \approx \frac{f_n(x)}{f_{n+1}(x)}, \quad |\lambda_1| \approx \frac{\|f_n\|}{\|f_{n+1}\|}, \quad \varphi_1(x) \approx \frac{f_n(x)}{\|f_n\|}, \text{ etc.}$$

For an arbitrary function $f(x)$ the inequality $f_1 \neq 0$ may be violated, and therefore to verify the result it is advisable to take another initial function and repeat the calculations. (Let the reader think of the ways in which we can recognize and investigate the cases when $\lambda_2 = \pm \lambda_1$.)

For the resolvent kernel (see (2.45)) we obtain the formula

$$\Gamma(x, \xi, \lambda)=\sum_{n=1}^{\infty} \lambda^{n-1} \sum_{j} \frac{1}{\lambda_j^n} \varphi_j(x) \varphi_j(\xi)$$

Interchanging the order of summation and taking advantage of the relation

$$\sum_{n=1}^{\infty} \lambda^{n-1} \frac{1}{\lambda_j^n}=\frac{1}{\lambda_j}\left(1-\frac{\lambda}{\lambda_j}\right)^{-1}=\frac{1}{\lambda_j-\lambda} \quad (|\lambda|<|\lambda_j|)$$

we obtain

$$\Gamma(x, \xi, \lambda)=\sum_{j} \frac{1}{\lambda_j-\lambda} \varphi_j(x) \varphi_j(\xi) \tag{14}$$

The above consideration guarantees the validity of formula (14) for small $|\lambda|$. But both members of this equality are analytic functions of λ in the entire λ-plane except at the points $\lambda=\lambda_j$, and therefore, since they coincide for small $|\lambda|$, they coincide identically everywhere (Sec. II.2.5). This means that formula (14) gives a representation of the resolvent kernel for all λ. In particular, we see that the resolvent of a symmetric kernel can only have poles of the first order.

When deriving formula (8) we considered the kernel $K(x, \xi)$ as being known. But in some problems we are given a series

$$\sum_{j} \frac{1}{\lambda_j} \varphi_j(x) \varphi_j(\xi) \tag{15}$$

where $\varphi_1, \varphi_2, \ldots$ is a system of pairwise orthogonal normalized functions and $\frac{1}{\lambda_j}$ are some coefficients such that the series of their squares is convergent (see (9)). Then series (15) is convergent in the mean square to a function $K(x, \xi)$ satisfying condition (1.7). (Prove the assertion on the basis of the material of the first half of Sec. 2.6.) Besides, $\varphi_j(x)$ $(j=1, 2, \ldots)$ are eigenfunctions of the kernel K corresponding to the characteristic values λ_j, and the kernel K has no other linearly independent eigenfunctions (why?).

In particular, it follows from formulas (12) and (14) that the kernel $K_n(x, \xi)$ and the resolvent kernel $\Gamma(x, \xi, \lambda)$ have the same set (3) of eigenfunctions as the original kernel $K(x, \xi)$, these functions belonging, respectively, to the characteristic values λ_j^n and $\lambda_j-\lambda$.

Substituting (14) into (2.44) we obtain the formula

$$u(x)=f(x)+\sum_{j} \frac{\lambda}{\lambda_j-\lambda}\left(\int_a^b f(\xi) \varphi_j(\xi)\, d\xi\right) \varphi_j(x) \tag{16}$$

for the solution which is valid for all noncharacteristic values of λ. If λ is a characteristic value then we see that for the solution to exist it is necessary that $f(x)$ be orthogonal to all eigenfunctions associated with this value. By the way, this fact is a direct consequence of Fredholm's third theorem (why?). If this orthogonality condition is fulfilled, the coefficients in the eigenfunctions belonging to the characteristic value λ entering into formula (16) must be replaced by arbitrary constants while all the other coefficients are left unchanged.

Now suppose that we are given a function $F(x)$ which is representable in the form

$$F(x)=\int_a^b K(x,\ \xi)\, f(\xi)\, d\xi \tag{17}$$

where $f \in L_2[a, b]$. Such a function $F(x)$ is said to be **(sourcewise) representable with the aid of the kernel** K. Substituting expansion (8) into the right-hand side we obtain

$$F(x)=\sum_j \frac{1}{\lambda_j}\left(\int_a^b f(\xi)\, \varphi_j(\xi)\, d\xi\right)\varphi_j(x)$$

Hence, it follows that *a function sourcewise representable with the aid of a kernel can be expanded in a series with respect to the eigenfunctions of the kernel.* This proposition is known as the **Hilbert-Schmidt theorem.**

The Hilbert-Schmidt theorem can be given the following interpretation. Let us consider the integral operator $Au(x)=$ $=\int K(x, \xi)\, u(\xi)\, d(\xi)$. Its domain of definition is the space $L_2[a, b]$ and its range is the subspace of functions representable with the aid of the kernel. The Hilbert-Schmidt theorem thus means that the sequence of functions (3) forms a Euclidean basis of this subspace. If any function from $L_2[a, b]$ can be approximated with elements of this subspace to any degree of accuracy (such cases are possible), the system of functions (3) is complete in $L_2[a, b]$. This is a most widely used method of testing a system of functions for completeness.

All the functional series considered above are convergent in the sense of L_2, that is in the mean square. In courses on integral equations the reader can find some additional conditions under which these series are uniformly convergent.

The results of Secs. 1-3 can be generalized without essential changes to the other types of Fredholm's equations of the second kind with a real symmetric kernel mentioned in Sec. 1.2 (as well as to equations containing integrals with respect to measure). For equations with unknown functions dependent on n indepen-

dent variables it is more natural to number the characteristic values and eigenfunctions with n indices but this is inessential. If we consider a system of equations the kernel K is understood as a square functional matrix and symmetry condition (1) is replaced by the relation

$$K(x, \xi) \equiv [K(\xi, x)]^*$$

where for real functions the asterisk designates the operation of transposing a matrix and for complex functions (to which the given results are also applicable) the same operation with the additional passage to the conjugate complex values. In the latter case expansion (8) takes the form

$$K(x, \xi) = \sum_j \frac{1}{\lambda_j} \varphi_j(x) [\varphi_j(\xi)]^*$$

In the complex case all the other formulas considered here should also be changed accordingly.

4. Passing from Nonsymmetric Kernel to Symmetric Kernel. There are cases when in the original equation the kernel is not symmetric. However, it is sometimes possible to pass to an equation with a symmetric kernel with the help of certain transformations, which enables us to resort to the theory of equations with symmetric kernels. For instance, let us consider the equation

$$u(x) = \lambda \int_a^b K(x, \xi) r(x) s(\xi) u(\xi) d\xi + f(x) \tag{18}$$

where $K(x, \xi) \equiv K(\xi, x)$ and r and s are arbitrary positive functions. In the general case the kernel $K(x, \xi) r(x) s(\xi)$ of equation (18) is nonsymmetric. But the transformation to the new unknown function

$$u = \sqrt{\frac{r(x)}{s(x)}} v$$

yields the equivalent equation

$$v(x) = \lambda \int_a^b K(x, \xi) \sqrt{r(x) s(x) r(\xi) s(\xi)}\, v(\xi) d\xi + \sqrt{\frac{s(x)}{r(x)}} f(x)$$

whose kernel is symmetric.

There is an important class of kernels reducible to symmetric ones which are termed **antisymmetric** or **skew-symmetric** kernels (**anti-Hermitian** or **skew-Hermitian** in the complex case). They are determined by the relation

$$K(\xi, x) = -[K(x, \xi)]^*$$

In this case the kernel $\frac{1}{i}K(x, \xi)$ is Hermitian (why?), and therefore, rewriting equation (2.16) in the form

$$u(x) = (\lambda i)\int_a^b \left[\frac{1}{i}K(x, \xi)\right] u(\xi)\, d\xi + f(x)$$

we can extend to it all the results of the theory of equations with Hermitian kernels. In particular, it should be noted that when constructing characteristic values we obtain real values of λi and hence the values of λ are pure imaginary.

In some cases it is preferable to use another technique for passing from a nonsymmetric kernel to a symmetric one. Let us consider general equation (2.16) with a real nonsymmetric kernel. Denoting x by s, multiplying both members by $K(s, x)$, integrating with respect to s and then replacing s by ξ in the one-fold integrals we obtain the relation (check it up!)

$$\int_a^b [K(\xi, x) - \lambda K_t(x, \xi)]\, u(\xi)\, d\xi = \int_a^b K(\xi, x)\, f(\xi)\, d\xi \qquad (19)$$

where K_t (the "left-iterated kernel") is defined as

$$K_t(x, \xi) = \int_a^b K(s, x) K(s, \xi)\, ds$$

From (19) and (2.16) we see that $u(x)$ satisfies the equation

$$u(x) = \lambda \int_a^b [K(x, \xi) + K(\xi, x) - \lambda K_t(x, \xi)]\, u(\xi)\, d\xi + \\ + f(x) - \lambda \int_a^b K(\xi, x)\, f(\xi)\, d\xi \qquad (20)$$

with a symmetric kernel (which turns out to be dependent on λ). It can be similarly shown that the solutions of equations (2.17) (adjoint to (2.16)) satisfy the equation

$$v(x) = \mu \int_a^b [K(x, \xi) + K(\xi, x) - \mu K_r(x, \xi)]\, v(\xi)\, d\xi$$

with the symmetric kernel K_r (the "right-iterated kernel") determined by the formula

$$K_r(x, \xi) = \int_a^b K(x, s) K(\xi, s)\, ds$$

Let the reader prove that all the characteristic values of the kernels K_l and K_r are positive and that the formulas

$$\mu=\lambda,\ \psi(x)=\sqrt{\lambda}\int_a^b K(x,\ \xi)\,\varphi(\xi)\,d\xi,$$

$$\varphi(x)=\sqrt{\lambda}\int_a^b K(\xi,\ x)\,\psi(\xi)\,d\xi \tag{21}$$

establish a one-to-one correspondence between the eigenfunctions of these kernels and also show that

$$\int_a^b \varphi_j(x)\,\varphi_k(x)\,dx=\int_a^b \psi_j(x)\,\psi_k(x)\,dx=\delta_{jk}$$

With the aid of these eigenfunctions we can represent the original kernel in the form of the sum of the series

$$K(x,\ \xi)=\sum_j \frac{1}{\sqrt{\lambda_j}}\psi_j(x)\,\varphi_j(\xi) \tag{22}$$

where λ_j are the characteristic values of the kernel K_l (or of K_r, which is the same).

Indeed, using formulas (21) we obtain

$$\int_a^b\int_a^b\left[K(x,\ \xi)-\sum_j \frac{1}{\sqrt{\lambda_j}}\psi_j(x)\,\varphi_j(\xi)\right]^2 dx\,d\xi=$$

$$=\int_a^b\int_a^b K(x,\ \xi)\,K(x,\ \xi)\,dx\,d\xi-\sum_j\frac{1}{\lambda_j}$$

the right-hand side being equal to zero according to formula (10) applied to K_l.

From formula (22) we can draw corollaries similar to those in Sec. 3 even if the kernel K itself has no characteristic values (for instance, such are Volterra's kernels). For example, let us consider an analogue of the Hilbert-Schmidt theorem. To this end we shall suppose that the function $F(x)$ can be represented by formula (17) with the aid of the kernel K. Substituting expansion (22) into the right-hand side we receive

$$F(x)=\sum_j \frac{1}{\sqrt{\lambda_j}}\left(\int_a^b f(\xi)\,\varphi_j(\xi)\,d\xi\right)\psi_j(x)$$

and hence the function $F(x)$ can be expanded with respect to the eigenfunctions of the kernel K_r. Similarly, every function $F(x)$ sourcewise representable with the aid of the kernel K by

means of the formula

$$F(x)=\int_a^b K(\xi, x) f(\xi)\, d\xi$$

(for the symmetric case both representations of $F(x)$ coincide) can be expanded with respect to the eigenfunctions of the kernel K_l.

5. Extremal Property of Characteristic Values. Characteristic values of Fredholm's equation with a symmetric kernel possess extremal properties which are completely analogous to the properties of eigenvalues of a self-adjoint mapping of a finite-dimensional Euclidean space into itself (see Sec. IV.1.7 and also Sec. VI.2.8). For simplicity, we shall limit ourselves to real one-dimensional scalar equations though the same properties take place in the general case.

Let us consider a quadratic functional

$$I\{y\}=\int_a^b\int_a^b K(x, \xi)\, y(x)\, y(\xi)\, dx\, d\xi \tag{23}$$

with a symmetric kernel K and the unknown function $y(x)$ subject to the integral constraint

$$\int_a^b y^2(x)\, dx = 1 \tag{24}$$

Here there is a close analogy with Sec. IV.1.7, which is easily seen if we rewrite functional (23) in the form $\int\left[\int K(x, \xi)\times\right.$ $\left.\times y(\xi)\, d\xi \cdot y(x)\right] dx$ and note that condition (24) determines a unit sphere in the space $L_2[a, b]$. A new feature (besides the fact that the space under consideration is infinite-dimensional) is that in the general case the system of eigenfunctions of the kernel K is no longer complete in $L_2[a, b]$ (for a degenerate kernel it is sure to be incomplete but for a nondegenerate kernel it may not be complete either). But this proves inessential in this problem.

Substituting expansion (8) into (23) we obtain, after some simple transformations, the relation

$$I\{y\}=\sum_l \frac{1}{\lambda_l} y_l^2 \tag{25}$$

where

$$y_l=\int_a^b y(x)\, \varphi_l(x)\, dx$$

The functions φ_l being normalized, the quantities y_l are the coefficients of the expansion of the function $y(x)$ with respect to the

orthogonal system of functions (3) (cf. [107], Sec. XVII.21). If system (3) were complete we should have $y(x)=\sum_j y_j\varphi_j(x)$; but in the general case we can only guarantee the equality

$$y(x)=\sum_j y_j\varphi_j(x)+y_0(x) \tag{26}$$

where $y_0(x)$ is a function orthogonal to all the functions (3) (check it up!). The substitution of this expansion into (24) results in

$$\sum_j y_j^2+\int_a^b y_0^2(x)\,dx=1 \tag{27}$$

Let us first suppose that $\lambda_j>0$ for all $j=1, 2, 3, \ldots$ and $\lambda_1\leqslant\lambda_2\leqslant\lambda_3\leqslant\ldots$. Since (25) and (27) imply

$$I\{y\}=\frac{1}{\lambda_1}-\sum_j\left(\frac{1}{\lambda_1}-\frac{1}{\lambda_j}\right)y_j^2-\frac{1}{\lambda_1}\int_a^b y_0^2(x)\,dx$$

we see that

$$I\{y\}\leqslant\frac{1}{\lambda_1}$$

where the sign of equality only occurs if $y_0(x)\equiv 0$ and $y_j=0$ for $\lambda_j>\lambda_1$ or, in other words, only if $y(x)$ is a normalized eigenfunction corresponding to λ_1 (see (26)). If such a function is known we can take it as $\varphi_1(x)$ and impose the additional constraint

$$\int_a^b y(x)\varphi_1(x)\,dx=0 \tag{28}$$

which similarly results in the relation $I\{y\}\leqslant\frac{1}{\lambda_2}$ where the equality sign only stands if $y(x)$ is an eigenfunction corresponding to λ_2 (and orthogonal to $\varphi_1(x)$). Then we introduce the two additional constraints

$$\int_a^b y(x)\varphi_1(x)\,dx=0,\quad \int_a^b y(x)\varphi_2(x)\,dx=0$$

etc.

According to (25), the condition that the inequality $I\{y\}\geqslant 0$ holds for all $y(x)\in L_2[a, b]$ is necessary and sufficient for all the characteristic values to be positive. Functional (23) and the kernel K possessing this property are said to be **positive-semidefinite.** If this inequality is strict for all the functions $y(x)\not\equiv 0$, the functional and the kernel are called **positive-definite.** From (25)

and (26) it follows that for functional (23) to be positive-definite it is necessary and sufficient that the following two conditions be fulfilled: all $\lambda_j > 0$ and system (3) is complete (why?). (In particular, it should be noted that by virtue of formula (11) the kernel K_2 is always positive-semidefinite.)

If all $\lambda_j < 0$, we have an analogous situation: $\frac{1}{\lambda_1} \leqslant I\{y\} \leqslant 0$ where λ_1 is the characteristic value with the least modulus. If there are characteristic values λ_j of either sign the value of $I\{y\}$ lies between the reciprocal of the negative characteristic value with the least modulus and the reciprocal of the least positive characteristic value, the extreme values of $I\{y\}$ being attained on the corresponding eigenfunctions. If such a function $\varphi_1(x)$ is known, we can add the orthogonality condition (28) and thus reduce appropriately the range of variation of $I\{y\}$, etc. According to Sec. 2, instead of introducing condition (28) we can pass to kernel (5) and consider the extremum problem for the corresponding quadratic functional with only one constraint (24).

The application of variational methods to finding a function giving a stationary value to functional (23) under constraint (24) (see Sec. VI.1.12) leads to the eigenfunctions of the kernel K. For a nonsymmetric kernel K the same problem leads to the eigenfunctions of its symmetric part $\frac{1}{2}[K(x, \xi) + K(\xi, x)]$ (check it up!).

The above extremal property can be used for applying the Ritz general method (Sec. VI.4.1) to the approximate calculation of characteristic values and eigenfunctions of a symmetric kernel. This method reduces the problem to the analogous problem for a quadratic form involving a finite number of variables. It is not necessary to know in advance the signs of all λ_j's because if, in the calculation process, $I\{y\} > 0$ for at least one function $y(x)$, there are positive numbers among λ_j's, that is, the maximum problem for the functional $I\{y\}$ makes sense. A similar reasoning is applied if $I\{y\}$ can assume a negative value.

In equation (20) we had a kernel dependent on λ. Such kernels (which may not be symmetric) are encountered in some other problems, and therefore we shall briefly discuss their properties. Let a kernel $K(x, \xi, \lambda)$ be an analytic function of λ in a domain (G) of the complex plane. Since in equation (2.16) we can regard λ as constant, the first three Fredholm's theorems hold for such kernels. The analogy with degenerate kernels (which can be justified rigorously) shows that the Fredholm determinant $D(\lambda)$ (see Secs. 2.1 and 2.2) whose zeros are characteristic values of the kernel K is an analytic function of λ in the domain (G). This means that if $D(\lambda) \not\equiv 0$ these zeros form either a finite set or

a sequence approaching the boundary of the domain (G) (into which the point $\lambda=\infty$ is also included). In particular, this is the case when $0\in(G)$ since $\lambda=0$ cannot be a characteristic value. However, the case $D(\lambda)\equiv 0$ is also possible; then all the values $\lambda\in(G)$ are characteristic. (For example, verify that for the kernel $K(x,\ \xi)=\frac{1}{2\lambda}+x\xi$, $-1\leqslant x,\ \xi\leqslant 1$ all the values $\lambda\neq 0$ are characteristic and that to each of them there corresponds the eigenfunction $\varphi_1(x)=\frac{1}{\sqrt{2}}$, the value $\lambda=\frac{3}{2}$ possessing the additional eigenfunction $\varphi_2(x)=\sqrt{\frac{3}{2}}\,x$.)

6. Equations with Self-Adjoint Operators. The properties of integral equations with symmetric kernels are extended to general equations (2.67) with completely continuous self-adjoint operator A. A linear operator A mapping a real Hilbert space (H) into itself is called **self-adjoint** if it coincides with its adjoint, in other words (cf. (2.66)) if

$$(\mathbf{Ax},\ \mathbf{y})=(\mathbf{x},\ \mathbf{Ay}) \text{ for all } \mathbf{x},\ \mathbf{y}\in(H) \tag{29}$$

We shall express these properties in terms of the eigenvalues of the operator A. It is necessary to bear in mind that the eigenvalues are the reciprocals of the characteristic values and therefore there is no characteristic value corresponding to the zero eigenvalue. Besides, a characteristic value cannot be equal to zero.

First of all we readily prove that if the operator A is distinct from zero, i.e. $A(H)\neq 0$, it possesses at least one nonzero eigenvalue. Let us denote this eigenvalue by λ_1 and the corresponding eigenvector (which is supposed to be normalized) by φ_1. The operator $A_1=\lambda_1(\cdot,\ \varphi_1)\varphi_1$ (in this notation the dot indicates the place where a function acted upon by the operator is substituted, that is $A_1\mathbf{x}=\lambda_1(\mathbf{x},\ \varphi_1)\varphi_1$) is self-adjoint and has only one nonzero eigenvalue λ_1 to which there corresponds only one eigenvector φ_1 (check it up!). Reasoning in the same way as in Sec. 2 we can easily show that if to the eigenvalue λ_1 of the operator A there corresponds only one eigenvector φ_1 the operator

$$A-\lambda_1(\cdot,\ \varphi_1)\varphi_1$$

has the same eigenvalues and eigenvectors as the operator A except λ_1 and φ_1. Proceeding in this manner we represent the operator A in the form of the sum

$$A=\sum_j \lambda_j(\cdot,\ \varphi_j)\varphi_j \tag{30}$$

extending over all the nonzero eigenvalues of the operator A and the corresponding eigenvectors. If there are k linearly independent

eigenvectors associated with an eigenvalue, the corresponding eigenvalue is repeated k times and the eigenvectors associated with it are chosen in such a way that they are orthogonal to each other and normalized (the eigenvectors corresponding to different eigenvalues are always orthogonal to each other).

Formula (30) implies the same consequences as those formulated in Sec. 3 for an integral operator. For instance, we have

$$A^2\mathbf{x} = A(A\mathbf{x}) = \sum_j \lambda_j (A\mathbf{x},\ \varphi_j)\varphi_j = \sum_j \lambda_j \left(\sum_k \lambda_k (\mathbf{x},\ \varphi_k)\varphi_k,\ \varphi_j\right)\varphi_j =$$
$$= \sum_{j,k} \lambda_j \lambda_k (\mathbf{x},\ \varphi_k)(\varphi_k,\ \varphi_j)\varphi_j = \sum_j \lambda_j^2 (\mathbf{x},\ \varphi_j)\varphi_j$$

(check it up!) which shows that

$$A^2 = \sum_j \lambda_j^2 (\cdot,\ \varphi_j)\varphi_j$$

and, generally,

$$A^n = \sum_j \lambda_j^n (\cdot,\ \varphi_j)\varphi_j \quad (n = 1,\ 2,\ 3,\ \ldots)$$

Multiplying these equalities by arbitrary coefficients and summing with respect to n we obtain

$$F(A) = \sum_j F(\lambda_j)(\cdot,\ \varphi_j)\varphi_j \tag{31}$$

where $F(\lambda)$ is an arbitrary one-valued analytic function such that $F(0) = 0$, and λ_j's are not its singular points. It follows from formula (30) that the norm of the operator A (see Sec. 1.3) is equal to $\max_j |\lambda_j|$ (according to the fourth Fredholm theorem, if there is an infinite number of eigenvalues they form a sequence convergent to zero).

Conversely, for any bounded sequence λ_j and any orthogonal system of normalized vectors φ_j formula (30) determines a bounded self-adjoint linear operator with eigenvalues λ_j and the corresponding eigenvectors φ_j (the case of a zero eigenvalue will be discussed later on). This operator is completely continuous if and only if $\lambda_j \underset{j \to \infty}{\longrightarrow} 0$. If vectors φ_j are fixed and the values of λ_j are arbitrarily varied we obtain a family of operators with given eigenvectors. These operators are pairwise commuting because

$$\left(\sum_j \lambda_j (\cdot,\ \varphi_j)\varphi_j\right)\left(\sum_j \mu_j (\cdot,\ \varphi_j)\varphi_j\right) = \sum_j \lambda_j \mu_j (\cdot,\ \varphi_j)\varphi_j$$

(In contrast to it, if the vectors φ and ψ are linearly independent and are not orthogonal, the operators $(\cdot,\ \varphi)\varphi$ and $(\cdot,\ \psi)\psi$ are not commuting. Prove it!) Now the meaning of the left-hand member of formula (31) becomes quite clear: if the norm of the operator A is sufficiently small (less than the radius of conver-

gence of the series for $F(\lambda))\,F(A)$ is the sum of the corresponding Maclaurent series; if the norm of A is not small, the right-hand member of (31) serves as the analytic continuation of this sum with respect to all λ_j.

In particular, formula (31) yields the expansion of the resolvent operator $\Gamma(\lambda)$ which is defined as

$$\Gamma(\lambda) = A(I-\lambda A)^{-1} = \sum_j \frac{\lambda_j}{1-\lambda\lambda_j}(\cdot,\ \varphi_j)\,\varphi_j$$

It can be easily proved that

$$(I+\lambda\Gamma(\lambda))\,(I-\lambda A) = I$$

(check it up!). Therefore, writing equation (2.67) in the form $(I-\lambda A)\,\mathbf{u} = \mathbf{f}$ and multiplying both sides by $I+\lambda\Gamma(\lambda)$ we obtain

$$\mathbf{u} = (I+\lambda\Gamma(\lambda))\,\mathbf{f}$$

This is nothing but formula (2.7) (or (2.44), which is the same).

We see that a more general approach clarifies the situation and leads to a better understanding of basic interrelations.

It follows from Sec. IV.1.4 that (H) is the orthogonal direct sum of its subspaces $A(H)$ and $A^{-1}(0)$ (in Sec. IV.1.4 we dealt with finite-dimensional spaces but this property was not used in the proof). In accordance with formula (30) the first subspace consists of the elements which can be expanded with respect to φ_j. The second subspace (if it is not a zero space) consists of all the eigenvectors corresponding to the zero eigenvalue (why?). Thus, *a system of orthogonal vectors φ_j is complete in (H) if and only if the operator* A *has no zero eigenvalue.* For equation (2.16) with a symmetric kernel this condition means that the homogeneous equation

$$\int_a^b K(x,\ \xi)\,u(\xi)\,d\xi \equiv 0 \qquad (a \leqslant x \leqslant b)$$

does not have a nontrivial solution.

All the above properties hold for an arbitrary complex Hilbert space (H) as well. In this case an operator fulfilling condition (29) is called **Hermitian** (cf. Sec. 1.2). In addition to the properties enumerated above we can mention that all the eigenvalues of a Hermitian operator are real. This is also true for a self-adjoint operator A in a real space (H) but we can speak about imaginary eigenvalues of the operator A in a real space only after the *complexification* of the space, that is after we have introduced the totality of the elements of the form $\mathbf{x}+i\mathbf{y}$ where

$$\mathbf{x} \in (H),\ \mathbf{y} \in (H),\ (\mathbf{x}_1+i\mathbf{y}_1,\ \mathbf{x}_2+i\mathbf{y}_2) =$$
$$= (\mathbf{x}_1,\ \mathbf{x}_2)+i\,(\mathbf{y}_1,\ \mathbf{x}_2)-i\,(\mathbf{x}_1,\ \mathbf{y}_2)+(\mathbf{y}_1,\ \mathbf{y}_2)$$

and extended the operator by means of the relation

$$A(x+iy)=Ax+iAy$$

Then the self-adjoint operator becomes Hermitian; check it up!

As in Sec. 4, we can introduce the notion of an **anti-Hermitian** operator for which $A^*=-A$, i.e. $(Ax, y)\equiv-(x, Ay)$. The properties of anti-Hermitian operators are the same as those of Hermitian operators but all the eigenvalues are pure imaginary. An arbitrary operator A can be represented in the form of a sum of a Hermitian operator and an anti-Hermitian operator by the formula

$$A=\frac{A+A^*}{2}+\frac{A-A^*}{2}$$

Dividing and multiplying the second summand by i we obtain another representation of an arbitrary operator in the form

$$A=A_1+iA_2$$

where A_1 and A_2 are Hermitian operators, the adjoint operator being expressed as $A^*=A_1-iA_2$ (why?).

It follows from (30) that

$$(Ax, x)=\sum_j \lambda_j |(x, \varphi_j)|^2$$

On the other hand, taking the scalar squares of both members of the equality

$$x=\sum_j (x, \varphi_j)\varphi_j+x_0$$

in which x_0 is orthogonal to all φ_j, we obtain the relation

$$\sum_j |(x, \varphi_j)|^2 \leqslant \|x\|^2$$

(known as **Bessel's inequality**; cf [107], Sec. XVII.27). This inequality turns into the equality if $x_0=0$, i.e. if x can be expanded with respect to the vectors φ_j. This makes it possible to derive a variational principle, analogous to that described in Sec. 5, for finding the eigenvalues λ_j, but we leave this to the reader.

§ 4. SPECIAL CLASSES OF INTEGRAL EQUATIONS

1. Volterra's Equations of the First Kind. Let us consider the equation

$$\int_a^x K(x, \xi)\, u(\xi)\, d\xi = f(x) \qquad (1)$$

It is clear that the condition

$$f(a)=0 \tag{2}$$

is necessary for its solvability. We shall suppose that the kernel $K(x, \xi)$ which is considered for $a \leqslant \xi \leqslant x$ is continuous on the diagonal $\xi = x$. Differentiating both sides of (1) with respect to x we arrive at the equation

$$K(x, x)\, u(x) + \int_a^x K'_x(x, \xi)\, u(\xi)\, d\xi = f'(x) \tag{3}$$

which is equivalent to (1) if condition (2) is fulfilled. If $K(x, x) \neq 0$ we can divide both members of (3) by $K(x, x)$ and thus obtain Volterra's equation of the second kind. On condition that $K(x, x) \equiv 0$ the above procedure can be carried out repeatedly.

Now take a typical example of an equation whose kernel turns into infinity on the diagonal:

$$\int_a^x \frac{H(x, \xi)}{(x-\xi)^\alpha} u(\xi)\, d\xi = f(x) \tag{4}$$

It is supposed that $0 < \alpha < 1$ and that the function $H(x, \xi)$ is finite on the diagonal. Let us replace x by s, multiply both sides by $\frac{1}{(x-s)^{1-\alpha}}$, integrate with respect to s from a to x and reverse the order of integration in the left member of the resultant relation. We then obtain (check it up!)

$$\int_a^x \left[\int_\xi^x \frac{H(s, \xi)}{(s-\xi)^\alpha (x-s)^{1-\alpha}} ds \right] u(\xi)\, d\xi = \int_a^x \frac{(s)}{(x-s)^{1-\alpha}} ds \tag{5}$$

It is also possible to pass from (5) to (4) and show that these equations are equivalent, but we shall not dwell on this question here.

Making in the expression in the square brackets on the left-hand side of equation (5) (this expression is interpreted as a kernel) the change of the variable of integration according to the formula $s = \xi + (x-\xi)t$ we see that this kernel is equal to

$$\int_0^1 \frac{H(\xi + (x-\xi)t, \xi)}{t^\alpha (1-t)^{1-\alpha}} dt$$

Therefore, on the diagonal, i.e. for $\xi = x$, it assumes the finite value

$$H(\xi, \xi) \int_0^1 t^{-\alpha} (1-t)^{\alpha-1} dt = H(\xi, \xi) \frac{\Gamma(\alpha)\Gamma(1-\alpha)}{\Gamma(1)} = \frac{\pi}{\sin \pi\alpha} H(\xi, \xi)$$

(e.g. see [107], Sec. XIV.18 and formula (II.3.27) in the present book). This enables us to apply to equation (5) the differentiation method described above.

For instance, let $H(x, \xi) \equiv 1$. Then equation (4) takes the form

$$\int_a^x \frac{1}{(x-\xi)^\alpha} u(\xi)\, d\xi = f(x) \tag{6}$$

In this case the kernel of the equation of type (5) is equal to $\frac{\pi}{\sin \pi\alpha}$, and hence, differentiating both sides, we obtain the solution

$$u(x) = \frac{\sin \pi\alpha}{\pi} \frac{d}{dx} \int_a^x \frac{f(s)}{(x-s)^{1-\alpha}}\, ds = |x-s = p-a| =$$

$$= \frac{\sin \pi\alpha}{\pi} \frac{d}{dx} \int_a^x \frac{f(a+x-p)}{(p-a)^{1-\alpha}}\, dp =$$

$$= \frac{\sin \pi\alpha}{\pi} \left[\frac{f(a)}{(x-a)^{1-\alpha}} + \int_a^x \frac{f'(a+x-p)}{(p-a)^{1-\alpha}}\, dp \right] \tag{7}$$

This relation can be regarded as the inversion formula for integral transformation (6).

Let us demonstrate the application of the result obtained to the solution of the following problem: a material point M slides without friction along a line L in a homogeneous field of gravitation (Fig. 115); it is required to determine the shape of the line for which the period of oscillations does not depend on the amplitude.

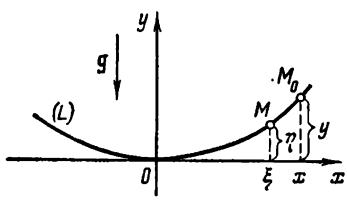


Fig. 115

To solve the problem let us assume that M_0 is the uppermost position and (0, 0) is the lowermost position of the oscillating point. Then, according to formula (VI.1.4), the period T of oscillations is equal to

$$T = 4 \int_0^x \sqrt{\frac{1+\eta'^2}{2g(y-\eta)}}\, d\xi = \frac{4}{\sqrt{2g}} \int_0^y \frac{1}{\sqrt{y-\eta}} \frac{\sqrt{1+\eta'^2}}{\eta'}\, d\eta$$

Putting $\frac{\sqrt{1+\eta'^2}}{\eta'} = u(\eta)$ we can apply inversion formula (7) with $\alpha = \frac{1}{2}$, which yields

$$\frac{\sqrt{1+y'^2}}{y'} = \frac{1}{\pi} \frac{T\sqrt{2g}}{4} \frac{1}{\sqrt{y}}$$

We have thus obtained a differential equation for $y(x)$. It is convenient to solve it by means of the substitution

$$y = R(1-\cos\varphi) \text{ where } R = \frac{gT^2}{16\pi^2}$$

The computations which we leave to the reader show that

$$x = R(\varphi + \sin\varphi)$$

Hence, the sought-for curve is a cycloid with vertex at the origin (check it up!). Equation (6) (sometimes referred to as **Abel's equation**) was investigated for $\alpha = \frac{1}{2}$ by N.H. Abel.

We can similarly treat many-dimensional analogues of Abel's equations. For instance, performing in the equation

$$\iint\limits_{(\sigma_{x,y})} [(y-\eta)^2-(x-\xi)^2]^{-\frac{1}{2}} u(\xi, \eta)\, d\xi\, d\eta = f(x, y)$$

where $(\sigma_{x,y})$ is a triangle $0 \leqslant \eta \leqslant y$, $|x-\xi| \leqslant y-\eta$, a transformation analogous to the above we arrive at the equality

$$\iint\limits_{(\sigma_{x,y})} u(\xi, \eta)\, d\xi\, d\eta = \frac{1}{\pi^2} \iint\limits_{(\sigma_{x,y})} [(y-\eta)^2-(x-\xi)^2]^{-\frac{1}{2}} f(\xi, \eta)\, d\xi\, d\eta$$

whence, by means of the differentiation, we obtain the inversion formula

$$u(x, y) = \frac{1}{2\pi^2}\left(\frac{\partial^2}{\partial y^2} - \frac{\partial^2}{\partial x^2}\right) \iint\limits_{(\sigma_{x,y})} [(y-\eta)^2-(x-\xi)^2]^{-\frac{1}{2}} f(\xi, \eta)\, d\xi\, d\eta$$

On such equations see [99], § I.6.

2. **Fredholm's Equations of the First Kind with Symmetric Kernel.** Let us consider the equation

$$\int_a^b K(x, \xi)\, u(\xi)\, d\xi = f(x) \tag{8}$$

with a real symmetric (or complex Hermitian) kernel and with a system of characteristic values (3.2) and eigenfunctions (3.3). Then, by the Hilbert-Schmidt theorem (Sec. 3.3), for the existence of the solution it is necessary that the function $f(x)$ be representable as an expansion with respect to the eigenfunctions (3):

$$f(x) = \sum_j (f, \varphi_j)\, \varphi_j(x) \tag{9}$$

If this condition is fulfilled the solution can be found in the form

$$u(x) = \sum_j u_j \varphi_j(x) \tag{10}$$

The substitution of (10) into (8) and the comparison with (9) yield

$$\frac{u_j}{\lambda_j}=(f,\ \varphi_j),\ \text{whence}\ u_j=\lambda_j(f,\ \varphi_j)$$

Therefore, for the solution to belong to $L_2[a,\ b]$ we should impose the condition that the function f satisfies the additional requirement $\sum_j \lambda_j^2 |(f,\ \varphi_j)|^2 < \infty$. This follows from the **Riesz-Fischer theorem** according to which for any orthogonal system of normalized functions $\varphi_j(x)$ series (10) converges in the mean square if and only if the condition $\sum_j |u_j|^2 < \infty$ holds. The sufficiency of the condition follows from the first definition of completeness for $L_2[a,\ b]$ (see Sec. 2.6). The necessity follows from the fact that the remainder of the last series is equal to the square of the mean square deviation between the partial sum of series (10) and its sum (why?).

As usual, to obtain the general solution of equation (8) it is necessary to add to its particular solution (10) the general solution of the corresponding homogeneous equation or, in other words, as was shown in Sec. 3.5, the general expression of the function orthogonal to all functions (3.3). Therefore, if system (3.3) is complete, the solution of equation (8) is unique; if otherwise, it is not unique and may even contain an infinite number of arbitrary constants. For instance, this is the case for any degenerate kernel (why?).

If all $\lambda_j > 0$ and $f(x)$ is such that equation (8) is solvable (but the functions $\varphi_j(x)$ are not given!) solution (10) can be obtained by the iteration method as the limit of the sequence of functions $u_n(x)$ constructed according to the formula

$$u_n(x)=u_{n-1}(x)+\lambda\left[f(x)-\int_a^b K(x,\ \xi)\,u_{n-1}(\xi)\,d\xi\right] \qquad (11)$$

$$(n=1,\ 2,\ 3,\ \ldots)$$

where $u_0(x)$ is an arbitrary function and $\lambda > 0$ is sufficiently small. Indeed, putting $u_n(x)=\sum_j u_{n,j}\varphi_j(x)$, $f(x)=\sum_j f_j\varphi_j(x)$ (the component of $u_0(x)$ orthogonal to the kernel disappears after the first iteration and therefore is inessential) we obtain from (11) the relation

$$u_{n,j}-u_{n-1,j}+\lambda\left(f_j-\frac{u_{n-1,j}}{\lambda_j}\right)=\left(1-\frac{\lambda}{\lambda_j}\right)u_{n-1,j}+\lambda f_j$$

It follows that for large n we approximately have

$$\Delta u_{n,j}=\left(1-\frac{\lambda}{\lambda_j}\right)\Delta u_{n-1,j}$$

and hence the process converges for $n \to \infty$ if $\left|1-\frac{\lambda}{\lambda_j}\right|<1$, i.e. if $0<\lambda<2\lambda_1$ where λ_1 is the least characteristic value. (In real problems λ_1 is usually unknown, and then the value of λ can be selected empirically by setting different values and investigating the corresponding rate of convergence.) Passing to the limit in (11) for $n \to \infty$ we see that the limit function satisfies equation (8) and therefore it is the sought-for solution.

It should be noted that the situation is in fact more complicated than it may seem from the preceding paragraph. Usually the rate of convergence is sufficiently high for the coefficients in the first eigenfunctions while for the subsequent eigenfunctions the convergence is worse due to the fact that the coefficient $1-\frac{\lambda}{\lambda_j}$ becomes close to unity. The whole process does not converge well if $u_0(x)$ differs considerably from the sought-for solution in its higher components. To improve the situation we can apply the iteration method beginning with different zero approximations $u_0(x)$; if after a number of iterations the results obtained are close to each other we have good reason to expect that the approximate solution we have found is sufficiently accurate.

3. Improperly Posed Problems. In connection with equation (8) it should be stressed that the errors occurring in the determination of the function f and in the computations (the round-off errors, the errors of the formulas of numerical integration, etc.) result in a term in $u_n(x)$ which is orthogonal to the kernel. If functions (3.3) do not form a complete system in $L_2[a, b]$ this term may grow as n increases. Such problems for which the accuracy (and even the existence of solutions) can be essentially affected by small variations of the initial data are said to be **improperly posed** ("not-well-posed"); naturally, the computational errors also affect the solutions of such problems. J.S. Hadamard (1865-1963), a noted French mathematician, was the first to describe some properties of improperly posed problems, but their systematic investigation was initiated by Academician A.N. Tikhonov (born in 1906), a prominent Soviet mathematician. There are many classical problems (e.g. the problem of numerical differentiation or the above problem of solving Fredholm's integral equation of the first kind) and also many important problems of modern mathematics which are improperly posed. Modern theoretical and applied mathematics offers a number of approaches for solving such problems which are referred to as **regularization** methods and make it possible to approximate solutions of improperly posed problems with solutions of certain well-posed problems.

Here we do not consider the regularization methods in detail and only mention, as an example, the technique based on adding

a term of the form $\alpha u(x)$ to equation (8) which makes it possible to pass to an equation of the second kind. The coefficient α should not be too large (because, if otherwise, this may essentially change the solution) or too small (so that it could provide a sufficient change in the character of the problem). Practically, an appropriate value of α is selected empirically by analysing some simpler problems whose solutions are known.

More precisely, in a real problem of that kind we usually have some information about the behaviour of the sought-for solution $u(x)$. Suppose that we manage to construct a function $\tilde{u}(x)$ whose behaviour imitates that of $u(x)$ (of course, the estimation of the quality of this "imitation" essentially depends on one's intuition and experience). Then we can compute the expression

$$\int_a^b K(x, \xi)\,\tilde{u}(\xi)\,d\xi = \tilde{f}(x)$$

and regard this relation as an integral equation in which $\tilde{u}(x)$ is the unknown function and the function $\tilde{f}(x)$ is given. Applying the numerical method, which we are going to test, to this equation we obtain an approximate solution which is then compared with the exact solution $\tilde{u}(x)$. If these solutions are sufficiently close we have good reason to expect that the method is applicable to the original equation (8) as well. If the information about the properties of the solution is insufficient we can try to test the method by taking several functions of the type of $\tilde{u}(x)$.

These general considerations are applicable to a considerably wider range of problems than the class of integral equation problems.

4. Fredholm's Equations of the First Kind. General Case. Let us consider equation (8) without assuming the kernel to be symmetric. We shall suppose that there exists a solution of the equation.

By analogy with Sec. 2 we can formulate the theoretical conditions for the existence and uniqueness of the solution on the basis of expansion (3.22) (it is supposed that the function $f(x)$ can be expanded with respect to the system ψ_j and the solution $u(x)$ with respect to the system φ_j).

Let us look for the solution of equation (8) in the form

$$u(x) = \sum_j u_j g_j(x)\,\nu(x) \tag{12}$$

where $g_1, g_2, g_3, \ldots$ is a complete system of functions on the interval $a \leqslant x \leqslant b$ and $\nu(x)$ is a "weight function". It is desirable to choose this weight function in such a way that it is close to the sought-for solution since this accelerates the convergence of

series (12). If little is known about the solution we can put $\nu(x) \equiv 1$. The substitution of (12) into (8) yields the equality

$$f(x) = \sum_j u_j h_j(x) \tag{13}$$

where

$$h_j(x) = \int_a^b K(x, \xi) g_j(\xi) \nu(\xi) d\xi \tag{14}$$

Thus, it remains to determine the coefficients u_j of expansion (13).

There are cases when expansion (13) is quite simple. For instance, this is the case if $h_j(x) = c_j x^j$ $(j = 0, 1, 2, \ldots)$ or if the functions h_j are pairwise orthogonal. If otherwise, we can obtain an orthogonal system of functions of the form

$$\chi_1 = h_1, \quad \chi_2 = h_2 - \frac{(h_2, \chi_1)}{(\chi_1, \chi_1)} \chi_1, \qquad \chi_3 = h_3 - \frac{(h_3, \chi_2)}{(\chi_2, \chi_2)} \chi_2 - \frac{(h_3, \chi_1)}{(\chi_1, \chi_1)} \chi_1, \quad \ldots \tag{15}$$

with the help of the orthogonalization process (e.g. see [107], Secs. VII.21 and XVII.20), expand f with respect to the system χ_j and then substitute into this expansion the expressions of the functions χ_j in terms of h_k $(k = 1, 2, \ldots, j)$ which are found from (15). We can also expand the functions f and h_j with respect to an arbitrary complete system of functions $\omega_k(x)$:

$$f(x) = \sum_k f_k \omega_k(x), \quad h_j(x) = \sum_k h_{jk} \omega_k(x) \qquad (j = 1, 2, 3, \ldots)$$

If these expressions are effectively found we then reverse the second expansion (i.e. determine the inverse matrix $(h_{jk})^{-1}$), express ω_k in terms of h_j and substitute the resultant expressions into the first expansion.

If the expansion of the function $h_j(x)$ in a power series begins with the term involving x^j $(j = 0, 1, 2, \ldots)$ we can make use of the equalities following from (13):

$$f(0) = u_0 h_0(0), \qquad f'(0) = u_0 h_0'(0) + u_1 h_1'(0),$$

$$f''(0) = u_0 h_0''(0) + u_1 h_1''(0) + u_2 h_2''(0)$$

etc.

For example, let us consider the equation

$$\int_0^\pi e^{ix \cos \xi} u(\xi) d\xi = f(x)$$

encountered in the theory of propagation of waves. We shall seek the solution in the form

$$u(x) = \sum_{n=0}^{\infty} u_n \cos nx$$

Then, by formula (14),

$$\hat{h}_n(x) = \int_0^{\pi} e^{ix\cos\xi}\cos n\xi\, d\xi = \pi i^n J_n(x)$$

(e.g. see [107], (XVII.113)). Expansion (13) takes the form

$$f(x) = \pi\sum_{n=0}^{\infty} u_n i^n J_n(x) = \pi\sum_{n=0}^{\infty} u_n i^n \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,(n+k)!}\left(\frac{x}{2}\right)^{n+2k}$$

(e.g. see [107], (XV.170)). It follows that

$$f(0) = \pi u_0,\ f'(0) = \frac{\pi i}{2}u_1,\ f''(0) = -\frac{\pi}{2}u_0 - \frac{\pi}{4}u_2,\ \text{etc.}$$

(check it up!), which makes it possible to compute the coefficients u_n in succession.

If the kernel of equation (8) has a finite jump $\varphi(\xi)$ for $x=\xi$ then, differentiating just as in Sec. 1, we can pass from this equation to Fredholm's equation of the second kind. Indeed, we then have

$$\frac{\partial K}{\partial x} = \varphi(\xi)\,\delta(x-\xi) + L(x,\ \xi)$$

where L is the derivative of K with respect to x taken without the term containing the delta function. After differentiating both sides of (8) with respect to x we obtain

$$\int_a^b [\varphi(\xi)\,\delta(x-\xi) + L(x,\ \xi)]\,u(\xi)\,d\xi = f'(x)$$

i.e.

$$u(x) = -\int_a^b \frac{L(x,\ \xi)}{\varphi(x)}\,u(\xi)\,d\xi + \frac{f'(x)}{\varphi(x)}$$

the latter being an equation of the second kind. If the kernel of equation (8) is continuous on the diagonal but its derivatives of the first order have jumps we differentiate repeatedly, etc. In particular, such a situation arises when the kernel of equation (8) is a Green function (see Sec. 1).

5. Generating Functions. A function $\Phi(t,\ x)$ is said to be a **generating function** for the system of functions

$$g_0(x),\ g_1(x),\ g_2(x),\ \ldots \qquad (16)$$

if

$$\Phi(t,\ x) = \sum_{n=0}^{\infty} a_n g_n(x)\,t^n \qquad (17)$$

where a_n are some nonzero constants. In other words, this means that functions (16) are obtained by expanding $\Phi(t, x)$ in powers of t. Many important systems of functions can be easily investigated by considering the properties of their generating functions. It should be noted that one and the same system of functions (16) may have different generating functions because the coefficients a_n can be varied arbitrarily.

Below are some examples of generating functions which are given without proof.

Consider the expansion

$$(1-2tx+t^2)^{-\beta-\frac{1}{2}} = \frac{\sqrt{\pi}}{\Gamma\left(\beta+\frac{1}{2}\right)} \sum_{n=0}^{\infty} T_n^{\beta}(x)\, t^n$$

Here the functions $T_n^{\beta}(x)$ turn out to be polynomials of degree n. They are known as **Gegenbauer's polynomials**. For a fixed β they form an orthogonal complete system of functions with the *weight function* $(1-x^2)^{\beta}$ (e.g. see [107], Sec. XVII.29) on the interval $-1 \leqslant x \leqslant 1$. In particular, we have $T_n^0(x) = P_n(x)$ (**Legendre's polynomials**) and $\sqrt{\frac{\pi}{2}}\, n T_n^{-\frac{1}{2}}(x) = T_n(x)$ (**Chebyshev's polynomials**) (see [107], Sec. XVII.29).

Now take the expansion

$$e^{-\frac{xt}{1-t}}(1-t)^{-a-1} = \sum_{n=0}^{\infty} \frac{1}{\Gamma(n+a+1)} L_n^a(x)\, t^n$$

The functions $L_n^a(x)$ are also polynomials of degree n. They are called **Laguerre's polynomials** and for a fixed a they form an orthogonal complete system of functions with weight function $x^a e^{-x}$ on the interval $0 \leqslant x < \infty$. (Here we mean the completeness in the space $L_2[0, \infty]$ with the weight function $x^a e^{-x}$. This is Hilbert's space with the scalar product $(f, g) = \int_0^{\infty} f(x)\, g(x)\, x^a e^{-x}\, dx$.)

Finally, in the expansion

$$e^{2tx-t^2} = \sum_{n=0}^{\infty} \frac{1}{n!} H_n(x)\, t^n \tag{18}$$

the functions $H_n(x)$ are again polynomials of degree n which are called **Hermite's polynomials**. They form an orthogonal complete system of functions with weight function e^{-x^2} on the interval $-\infty < x < \infty$.

The polynomials enumerated here possess various useful properties which will not be discussed here.

Suppose that for a generating function (17) the system of functions (16) turns out to be real and orthogonal with weight function $\rho(x) \geqslant 0$ on the interval $a \leqslant x \leqslant b$. Let us consider the equation

$$\int_a^b \Phi(x, \xi) \rho(\xi) u(\xi) d\xi = f(x) \tag{19}$$

where the function $f(x)$ is analytic in the vicinity of the point $x=0$. Its solution can be sought in the form

$$u(x) = \sum_{j=0}^{\infty} u_j g_j(x) \tag{20}$$

The substitution of (17) and (20) into (19) yields, by the orthogonality conditions, the relation

$$\sum_{n=0}^{\infty} a_n u_n \| g_n \|^2 x^n = f(x)$$

where

$$\| g_n \|^2 = \int_a^b [g_n(x)]^2 \rho(x) dx$$

It follows that

$$u_n = \frac{f^{(n)}(0)}{n! a_n \| g_n \|^2} \qquad (n = 0, 1, 2, \ldots)$$

Substituting these values into (20) we obtain the sought-for solution which is unique if system (16) is complete.

For example, taking generating function (18) we can solve the equation

$$\int_{-\infty}^{\infty} e^{-(x-\xi)^2} u(\xi) d\xi = f(x) \tag{21}$$

(occurring in the theory of heat propagation) with the aid of a series in Hermite's polynomials. When computing the coefficients we should take into account the formula $\| H_n \|^2 = 2^n n! \sqrt{\pi}$ which we shall not prove here.

It should be noted that taking an initial kernel $K(x, \xi)$ for which we can solve equation (8) we can then construct many other kernels possessing the same property by applying the following general method. Let us consider the equation

$$\int_a^b L_x K(x, \xi) u(\xi) d\xi = f_1(x) \tag{22}$$

where L_x is a linear (for instance, differential) operator (the subscript x indicates that the operator is only applied to the variable x,

i.e. ξ is regarded as a constant parameter when the operator acts upon a function dependent on x and ξ). Suppose that we can find the general solution of the equation

$$L_x f(x) = f_1(x)$$

Then, substituting this solution into the right-hand side of (8), we thus reduce equation (22) to (8). Hence, we can construct the solution of equation (22) as well.

For instance, from expansion (18) which can be rewritten in the form

$$e^{-(x-\xi)^2} = e^{-\xi^2} \sum_{n=0}^{\infty} \frac{1}{n!} H_n(\xi) x^n$$

it follows that

$$\frac{d^k}{dx^k} e^{-(x-\xi)^2} = \left\{ \left[\frac{d^k}{dx^k} e^{-(x-\eta)^2} \right]_{x=0} \right\}_{\eta=\xi-x} =$$
$$= \{e^{-\eta^2} H_k(\eta)\}_{\eta=\xi-x} = (-1)^k e^{-(x-\xi)^2} H_k(x-\xi)$$

The last relation is due to the equality

$$\sum \frac{1}{n!} H_n(-x) t^n = e^{-2tx-t^2} = \sum \frac{1}{n!} H_n(x) (-t)^n$$

implying $H_n(-x) = (-1)^n H_n(x)$. (This is a typical example in which the properties of functions are established with the aid of their generating function.) Thus, we can solve any equation of the form

$$\int_{-\infty}^{\infty} e^{-(x-\xi)^2} \cdot \sum_{k=0}^{N} a_k H_k(x-\xi)\, u(\xi)\, d\xi = f_1(x) \qquad (23)$$

where

$$a_k = \text{const} \quad (k = 0, 1, 2, \ldots, N)$$

if the simple equation

$$\sum_{k=0}^{N} (-1)^k a_k \frac{d^k f}{dx^k} = f_1(x)$$

and then equation (21) have been solved beforehand. (By the way, equation (23) can be solved by means of the formula

$$e^{-(x-\xi)^2} H_k(x-\xi) = (-1)^k \frac{d^k}{dx^k} e^{-(x-\xi)^2} = (-1)^k e^{-\xi^2} \sum_{n=0}^{\infty} \frac{1}{n!} H_{n+k}(\xi) x^n$$

if we represent $u(\xi)$ as a series in Hermite's polynomial with undetermined coefficients.) Here we can also put $N = \infty$.

There is another approach to solving equation (22) applicable when the kernel, as it sometimes happens, satisfies an equation of the form

$$L_x K(x, \xi) = M_\xi K(x, \xi) \tag{24}$$

where M_ξ is a linear operator acting upon the variable ξ. Substituting (24) into (22) and passing, if it is possible, to the adjoint operator M_ξ^* acting on u (for instance, with the help of integration by parts) we arrive at the equation

$$\int K(x, \xi) M_\xi^* u(\xi) d\xi = f_2(x)$$

After this equation has been solved we must return from $M_\xi^* u(\xi)$ to $u(\xi)$.

The problems under consideration have much in common with the **problem of moments** which consists in constructing a function $u(x)$ by its **moments**

$$\int_a^b u(x) x^n \rho(x) dx = M_n \qquad (n = 0, 1, 2, \ldots) \tag{25}$$

where $\rho(x) \geqslant 0$ is a given weight function. It should be noted that the weight function can be varied because we can always rewrite equation (25) in the form

$$\int_a^b v(x) x^n \rho_1(x) dx = M_n \text{ where } v(x) = \frac{\rho(x)}{\rho_1(x)} u(x)$$

We can take advantage of this property to obtain a function $v(x)$ of a simpler structure (for instance, close to a constant) by using the information about $u(x)$. As the interval (a, b) we can take one of the intervals $(-1, 1)$, $(0, \infty)$ or $(-\infty, \infty)$ because any interval can be reduced to them by means of a linear change of the variable of integration (check it up!).

One of the methods for solving the problem consists in constructing a system of polynomials

$$Q_n(x) = \sum_{j=0}^{n} a_{nj} x^j \qquad (n = 0, 1, 2, \ldots)$$

orthogonal on the interval (a, b) with the weight function $\rho(x)$ by means of the orthogonalization process applied to the sequence of powers of x. If we seek the solution in the form

$$u(x) = \sum_{n=0}^{\infty} c_n Q_n(x) \tag{26}$$

the coefficients are expressed by the formulas

$$c_n = \frac{1}{\| Q_n \|^2} \int u(x)\, Q_n(x)\, \rho(x)\, dx = \frac{1}{\| Q_n \|^2} \sum_{j=0}^{n} a_{nj} M_j \quad (n = 0, 1, 2, \ldots)$$

Of course, for this technique to be applicable series (26) must be convergent. This method can also be used when the moments are taken with respect to some other system of functions instead of the sequence of powers.

There is another method based on passing to the integral equation of the first kind. To this end, we expand an appropriately chosen kernel $K(x, \xi)$ with respect to powers of ξ:

$$K(x, \xi) = \sum_{n=0}^{\infty} k_n(x)\, \xi^n$$

Relations (25) then yield the equation

$$\int_a^b K(x, \xi)\, \rho(\xi)\, u(\xi)\, d\xi = \sum_{n=0}^{\infty} M_n k_n(x)$$

to which any of the above methods can be applied.

6. Volterra's Equations with Difference Kernels. The equation

$$\alpha u(x) + \int_0^x k(x-\xi)\, u(\xi)\, d\xi = f(x) \qquad (\alpha = \text{const}) \tag{27}$$

with a difference kernel $k(x-\xi)$ (which is an equation of the second kind for $\alpha \neq 0$ and of the first kind for $\alpha = 0$) can be solved by means of Laplace's transformation as was described in Sec. III.2.4. Denoting with capital letters the Laplace transforms of the functions under consideration we obtain from (27) the equation

$$\alpha U(p) + K(p)\, U(p) = F(p) \tag{28}$$

whence

$$U(p) = \frac{F(p)}{\alpha + K(p)} \tag{29}$$

Now it only remains to apply the inversion formula (III.1.19). Since the functions $F(p)$ and $K(p)$ are analytic in the half-plane $\operatorname{Re} p > \text{const}$ and tend to zero for $p \to \infty$, the function $U(p)$ also possesses these properties guaranteeing the applicability of the inversion formula if $\alpha \neq 0$. If $\alpha = 0$, that is if the equation under consideration is of the first kind, we should additionally impose the condition that the right-hand member of (29) tends to zero for $p \to \infty$, $\operatorname{Re} p > \text{const}$.

If equation (27) is solved only on a finite interval $0 \leqslant x \leqslant x_0$ the functions $k(x)$ and $f(x)$ can be arbitrarily extended to the

exterior of this interval; for instance, they can be put equal to zero, after which Laplace's transformation is applied as above.

It can readily be shown that the iterated kernels (Sec. 2.9) are also difference kernels. For instance, the second iterated kernel is equal to

$$\int_{\xi}^{x} k(x-\eta)\,k(\eta-\xi)\,d\eta = \int_{0}^{x-\xi} k((x-\xi)-s)\,k(s)\,ds \quad (\eta-\xi=s)$$

Therefore, the resolvent kernel of the equation

$$u(x)=\lambda\int_{0}^{x} k(x-s)\,u(s)\,ds+f(x)$$

depends on $x-\xi$ (we shall designate it by $\gamma(x-\xi,\ \lambda)$. From (2.9) it follows that it satisfies the equation

$$\gamma(x-\xi,\ \lambda)=\lambda\int_{\xi}^{x} k(x-\eta)\,\gamma(\eta-\xi,\ \lambda)\,d\eta+k(x-\xi)$$

Putting $\xi=0$ we obtain, by analogy with (29), the Laplace transform of the resolvent kernel:

$$\Gamma(p,\ \lambda)=\frac{K(p)}{1-\lambda K(p)} \tag{30}$$

For instance, let us consider the equation

$$u(x)=\lambda\int_{0}^{x}(x-\xi)^{\alpha}\,u(\xi)\,d\xi+f(x) \qquad (\alpha>-1)$$

By formula (III.1.6) we have

$$K(p)=\Gamma(\alpha+1)\,p^{-(\alpha+1)}$$

whence

$$\Gamma(p,\ \lambda)=\frac{\Gamma(\alpha+1)\,p^{-(\alpha+1)}}{1-\lambda\Gamma(\alpha+1)\,p^{-(\alpha+1)}}$$

(pay attention to the different meanings of the letter Γ in the right-hand and left-hand sides of this formula!). Representing the fraction as the sum of the corresponding geometric series and using formula (III.1.6) we obtain an expression of the resolvent kernel in the form of the sum of a series (check it up!):

$$\gamma(x-\xi,\ \lambda)=\sum_{n=1}^{\infty}\lambda^{n-1}\frac{[\Gamma(\alpha+1)]^{n}}{\Gamma(n\alpha+n)}(x-\xi)^{n\alpha+n-1}$$

For $\alpha=0,\ 1,\ 2,\ \ldots$ this series can be written as a finite sum (e.g. see the method applied in [107] to series (XVII.52)).

If the kernel of equation (27) satisfies a homogeneous linear differential equation with constant coefficients with respect to x we can also apply the following technique. For instance, let

$$k'' + qk' + rk = 0 \tag{31}$$

Differentiating equation (27) we obtain

$$\alpha u'(x) + k(0)\, u(x) + \int_0^x k'(x-\xi)\, u(\xi)\, d\xi = f'(x) \tag{32}$$

and the repeated differentiation yields

$$\alpha u''(x) + k(0)\, u'(x) + k'(0)\, u(x) + \int_0^x k''(x-\xi)\, u(\xi)\, d\xi = f''(x) \tag{33}$$

Multiplying equation (27) by r and equation (32) by q and adding them to (33) we obtain, by formula (31), the equation

$$\alpha u'' + [k(0) + \alpha q]\, u' + [k'(0) + qk(0) + \alpha r]\, u = f''(x) + qf'(x) + rf(x)$$

After solving this simple differential equation (what initial conditions should be taken for $u(x)$?) we find $u(x)$. This technique is also applicable when $\alpha = 0$.

This method of passing to differential equation has a still simpler form for Fredholm's equations of the second kind if the kernel, say $K(x, \xi)$, satisfies, with respect to x, a homogeneous linear differential equation whose coefficients may also depend on x. (In the latter case, however, the solution may be connected with some difficulties.) It should be noted that such kernels are necessarily degenerate, and therefore the techniques considered in Sec. 2.1 are also applicable to them.

In connection with the method discussed above see [105].

7. Fredholm's One-Dimensional Equations with Difference Kernels. The Fredholm theory cannot be applied to the equation

$$u(x) = \lambda \int_{-\infty}^{\infty} r(x-\xi)\, u(\xi)\, d\xi + f(x) \tag{34}$$

because in this case the integral (1.7) is always divergent to infinity (why?). For this equation we usually use Fourier's transformation (e.g. see [107], § XVII.5)

$$\hat{f}(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x)\, e^{-ikx}\, dx \tag{35}$$

As is known, it can be applied to any function absolutely integrable on the interval $-\infty < x < \infty$, the inversion formula being

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(k)\, e^{ikx}\, dk \tag{36}$$

(e.g. see [107], Sec. XVII.32). Besides the well-known linearity property of Fourier's transformation we shall also use the following property of convolutions whose proof is analogous to the case of Laplace's transforms (see Sec. III.1.3): Fourier's transform of the function $\int_{-\infty}^{\infty} f_1(x-\xi) f_2(\xi)\, d\xi$ is $2\pi \hat{f}_1(k) \hat{f}_2(k)$.

Let us use these properties for solving equation (34) under the assumption that all the functions involved are absolutely integrable on the whole axis. By analogy with (29) and (30), we obtain Fourier's transforms of the solution and of the resolvent kernel which are, respectively,

$$\hat{u}(k) = \frac{\hat{f}(k)}{1 - 2\pi\lambda \hat{r}(k)}, \quad \hat{\gamma}(k, \lambda) = \frac{\hat{r}(k)}{1 - 2\pi\lambda \hat{r}(k)} \tag{37}$$

(it can easily be shown that in this case the resolvent kernel is also of the form $\gamma(x-\xi, \lambda)$) if the denominator has no zeros on the k-axis. Now we must take advantage of the inversion formula (36) and then transform, if possible, the resultant integral to a more convenient form by using the method of Sec. III.1.5. If $1 - 2\pi\lambda \hat{r}(k)$ has zeros for real k, then, generally speaking, equation (34) does not possess an absolutely integrable solution on the entire x-axis.

This method can also be applied to the equation of the first kind

$$\int_{-\infty}^{\infty} r(x-\xi)\, u(\xi)\, d\xi = f(x)$$

but in this case we must additionally require that the ratio $\frac{\hat{f}(k)}{\hat{r}(k)}$ should admit of the application of Fourier's inverse transformation, that is it should be absolutely integrable on the entire k-axis.

For an n-dimensional equation analogous to (34) we can similarly apply n-dimensional Fourier's transformation (Sec. III.3.3). For $n = 2$ such an analogue is written as

$$u(x, y) = \lambda \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} r(x-\xi, y-\eta)\, u(\xi, \eta)\, d\xi\, d\eta + f(x, y)$$

The uniqueness of the solution of equation (34) is only guaranteed in the class of absolutely integrable functions on the entire x-axis. For a wider class of functions homogeneous equation (34) may possess a nontrivial solution of the form $u(x)=e^{ax}$. The substitution of this expression into (34) and the change of variable $\xi \to x-\xi$ show that for such a solution to exist it is necessary and sufficient that

$$\lambda \int_{-\infty}^{\infty} r(\xi) e^{-a\xi} d\xi = 1 \tag{38}$$

(provided that the integral is convergent). It can also be shown that if a is an m-fold root of equation (38), the functions e^{ax}, xe^{ax}, ..., $x^{m-1}e^{ax}$ satisfy homogeneous equation (34) (e.g. see [107], Sec. XV.17).

The conditions that all the functions involved are absolutely integrable and that the denominator in (37) does not vanish are too restrictive and can be relaxed if, as was described in Sec. III.3.2, we pass to complex values of k. In this connection the following proposition is of use.

Lemma. *Let*

$$\begin{aligned} &|f(x)| \leqslant M_- e^{\alpha_- x} && \text{for } -\infty < x \leqslant 0, \\ &|f(x)| \leqslant M_+ e^{\alpha_+ x} && \text{for } 0 \leqslant x < \infty, \\ &|\varphi(x)| \leqslant N_- e^{\beta_- x} && \text{for } -\infty < x \leqslant 0, \\ &|\varphi(x)| \leqslant N_+ e^{\beta_+ x} && \text{for } 0 \leqslant x < \infty \end{aligned} \tag{39}$$

where $\beta_+ < \alpha_-$, $\alpha_+ < \beta_-$. *Then the convolution* $\int_{-\infty}^{\infty} f(x-\xi)\varphi(\xi)\,d\xi$ *exists and satisfies the inequalities*

$$\left| \int_{-\infty}^{\infty} f(x-\xi)\varphi(\xi)\,d\xi \right| \leqslant P_- e^{\gamma_- x} \quad \text{for} \quad -\infty < x \leqslant 0,$$

$$\left| \int_{-\infty}^{\infty} f(x-\xi)\varphi(\xi)\,d\xi \right| \leqslant P_+ e^{\gamma_+ x} \quad \text{for} \quad 0 \leqslant x < \infty$$

where γ_- *is the least of the numbers* α_- and β_-, γ_+ *is the greatest of the numbers* α_+ *and* β_+, *and* P_- *and* P_+ *are some constants.*

To prove the lemma we suppose that $x \geqslant 0$, $\alpha_+ \neq \beta_+$ (the investigation of all the other cases is left to the reader). Then

we have

$$\left|\int_{-\infty}^{\infty} f(x-\xi)\varphi(\xi)\,d\xi\right| \leqslant \int_{-\infty}^{0} |f(x-\xi)||\varphi(\xi)|\,d\xi + \int_{0}^{x} |f(x-\xi)||\varphi(\xi)|\,d\xi + \int_{x}^{\infty} |f(x-\xi)||\varphi(\xi)|\,d\xi \leqslant$$

$$\leqslant \int_{-\infty}^{0} M_+ e^{\alpha_+(x-\xi)} N_- e^{\beta_- \xi}\,d\xi + \int_{0}^{x} M_+ e^{\alpha_+(x-\xi)} N_+ e^{\beta_+ \xi}\,d\xi +$$

$$+ \int_{x}^{\infty} M_- e^{\alpha_-(x-\xi)} N_+ e^{\beta_+ \xi}\,d\xi = \frac{M_+ N_-}{\beta_- - \alpha_+} e^{\alpha_+ x} +$$

$$+ \frac{M_+ N_+}{\beta_+ - \alpha_+}(e^{\beta_+ x} - e^{\alpha_+ x}) + \frac{M_- N_+}{\alpha_- - \beta_+} e^{\beta_+ x}$$

which implies the validity of the assertion of the lemma.

Now let us suppose that the functions $f(x)$, $r(x)$ and $u(x)$ in equation (34) satisfy inequalities of form (39), the exponents being, respectively, $\alpha_\pm$, $\beta_\pm$ and $\gamma_\pm$. Besides let

$$\beta_+ < \beta_-,\ \alpha_+ < \beta_-,\ \gamma_+ < \beta_-,\ \beta_+ < \alpha_-,\ \beta_+ < \gamma_- \qquad (40)$$

Then we write $f(x) = f_+(x) + f_-(x)$, $u(x) = u_+(x) + u_-(x)$ (see Sec. III.3.2) and thus obtain from (34) the equality

$$u_+(x) - \lambda \int_{-\infty}^{\infty} r(x-\xi) u_+(\xi)\,d\xi - f_+(x) = -u_-(x) +$$

$$+ \lambda \int_{-\infty}^{\infty} r(x-\xi) u_-(\xi)\,d\xi + f_-(x) \qquad (41)$$

Let us denote the equal values of the left-hand and right-hand sides of (41) by $\varphi(x)$. Applying for $x < 0$ the above lemma to the left-hand side of (41) and for $x > 0$ to the right-hand side, we find that $\varphi(x)$ satisfies inequality (39) with the exponents $\beta_\pm$, and therefore, by virtue of Sec. III.3.2, the function $\hat{\varphi}(k)$ as well as the function $\hat{r}(k)$ are analytic in the strip

$$-\beta_- < \operatorname{Im} k < -\beta_+ \qquad (42)$$

Both functions tend to zero for $k \to \infty$ in this strip.

Passing to Fourier's inverse transforms in equality (41) we obtain

$$\hat{u}_+(k) - 2\pi\lambda \hat{r}(k)\hat{u}_+(k) - \hat{f}_+(k) = -\hat{u}_-(k) +$$
$$+ 2\pi\lambda \hat{r}(k)\hat{u}_-(k) + \hat{f}_-(k) = \hat{\varphi}(k)$$

whence

$$\hat{u}_+(k) = \frac{\hat{f}_+(k)+\hat{\varphi}(k)}{1-2\pi\lambda\hat{r}(k)}, \quad \hat{u}_-(k) = \frac{\hat{f}_-(k)-\hat{\varphi}(k)}{1-2\pi\lambda\hat{r}(k)}$$

Using the inversion formula (III.3.10) we receive

$$u(x) = \int_{-\infty+is_+}^{\infty+is_+} \frac{\hat{f}_+(k)+\hat{\varphi}(k)}{1-2\pi\lambda\hat{r}(k)} e^{ikx}\,dk + \int_{-\infty+is_-}^{\infty+is_-} \frac{\hat{f}_-(k)-i\hat{\varphi}(k)}{1-2\pi\lambda\hat{r}(k)} e^{ikx}\,dk \quad (43)$$

where $s_+ < -\gamma_+$, $s_- > -\gamma_-$. By the arbitrariness of s_+ and s_-, we can assume that on the contours of integration there are no zeros of the function $1-2\pi\lambda\hat{r}(k)$. (In the case of real k we do not have these arbitrary quantities at our disposal.)

If we additionally suppose that $s_+ < -\alpha_+$, $-\beta_- < s_+ < -\beta_+$ in (43) and impose similar conditions on s_- (here we make use of conditions (40)), each of the integrals in (43) can be split into two. Combining the integrals with $\hat{\varphi}(k)$ we can apply Cauchy's residue theorem to the resultant expression, which, by the arbitrariness of $\hat{\varphi}(k)$, yields the formula

$$u(x) = \int_{-\infty+is_+}^{\infty+is_+} \frac{\hat{f}_+(k)}{1-2\pi\lambda\hat{r}(k)} e^{ikx}\,dk + \int_{-\infty+is_-}^{\infty+is_-} \frac{\hat{f}_-(k)}{1-2\pi\lambda\hat{r}(k)} e^{ikx}\,dk + \sum_j A_j e^{ik_j x} \quad (44)$$

where A_j are arbitrary constants and the sum extends over all the zeros of the function $1-2\pi\lambda\hat{r}(k)$ (for the given λ) falling into the strip (42). In the case of an m-fold zero k_j the computation of the residue by formula (II.3.6) shows that the terms with $xe^{ik_j x}, \ldots, x^{m-1}e^{ik_j x}$ must be added to the last sum in (44) (why?). This situation is coherent with what was said above about non-trivial solutions of homogeneous equation (34) (if we put $a = ik$) because the last sum in (44) is the solution of equation (34) for $f(x) \equiv 0$. This sum can be interpreted as an eigenfunction of equation (34); we see that the characteristic values λ are no longer discrete (as in the Fredholm theory) but cover continuously the region which is the image of the strip (42) under the mapping $\lambda = (2\pi\hat{r}(k))^{-1}$ (why?).

As an example, let us consider the equation

$$u(x) = \lambda \int_{-\infty}^{\infty} e^{-|x-\xi|} u(\xi)\,d\xi + f(x) \quad (45)$$

Here we have $\beta_+ = -1$, $\beta_- = 1$; hence, by (40), the moduli of the nonhomogeneity term $f(x)$ and of the corresponding solution $u(x)$ can grow exponentially for $x \to \pm\infty$ but not faster than $e^{c|x|}$ where $c < 1$, that is $\alpha_+ = 1$, $\alpha_- = -1$. We have

$$\hat{r}(k) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ikx} e^{-|x|}\,dx = \frac{1}{2\pi}\int_{-\infty}^{0} e^{-ikx+x}\,dx + \frac{1}{2\pi}\int_{0}^{\infty} e^{-ikx-x}\,dx =$$

$$= \frac{1}{\pi(1+k^2)} \qquad (|\operatorname{Im} k| < 1)$$

(check it up!), and therefore for the first integral (44) the inverse image of the function

$$\frac{1}{1-2\pi\lambda\hat{r}(k)} = 1 + \frac{2\lambda}{k^2+1-2\lambda}$$

is

$$2\pi\delta(x) + \int_{-\infty+is_+}^{\infty+is_+} \frac{2\lambda}{k^2+1-2\lambda} e^{ikx}\,dk \tag{46}$$

The integrand function in (46) has poles at the points $k_{1,2} = \pm i\sqrt{1-2\lambda}$. Let λ be a number (complex, in the general case) such that $|\operatorname{Im} k_{1,2}| < 1$ and $k_{1,2} \neq 0$. (Let the reader investigate the other possible cases.) If $x > 0$ we can add the upper semicircle of (infinitely) large radius to the contour of integration in (46) and thus pass to a closed contour which, according to the choice of $s_+ < -\alpha_+$, encloses both poles. By Cauchy's residue theorem, we conclude that the integral in (46) is equal to

$$2\pi i\left(\frac{2\lambda}{2k_1} e^{ik_1x} + \frac{2\lambda}{2k_2} e^{ik_2x}\right) = -4\pi\lambda\,\frac{\sinh\sqrt{1-2\lambda}\,x}{\sqrt{1-2\lambda}}$$

Similarly, for $x < 0$, closing the contour of integration in (46) by the lower semicircle of large radius, we see that the integral is equal to zero.

The first integral in (44) is the preimage of a product and therefore it can be obtained as the convolution of the inverse images of the factors, that is it is equal to

$$\frac{1}{2\pi}\left[\int_{-\infty}^{\infty} f_+(x-\xi)\,2\pi\delta(\xi)\,d\xi - \int_{0}^{\infty} f_+(x-\xi)\,4\pi\lambda\,\frac{\sinh\sqrt{1-2\lambda}\xi}{\sqrt{1-2\lambda}}\,d\xi\right] =$$

$$= f_+(x) - \frac{2\lambda}{\sqrt{1-2\lambda}}\int_{0}^{x} f_+(x-\xi)\,\sinh\left(\sqrt{1-2\lambda}\xi\right)d\xi$$

The similar computation of the second integral in (44) yields the same result but f_+ must be replaced by f_- (check it up!). The-

refore, from (44) we obtain the solution of equation (45) (under the above conditions on λ) which has the form

$$u(x)=f(x)-\frac{2\lambda}{\sqrt{1-2\lambda}}\int_0^x \sinh\left[\sqrt{1-2\lambda}\,(x-\xi)\right]f(\xi)\,d\xi+$$
$$+A_1e^{-\sqrt{1-2\lambda}\,x}+A_2e^{\sqrt{1-2\lambda}\,x}$$

where A_1, A_2 are arbitrary constants.

It should be noted that equation (45) can be reduced to a differential equation; to this end we rewrite it in the form

$$u(x)=\lambda e^{-x}\int_{-\infty}^{x} e^{\xi}u(\xi)\,d\xi+\lambda e^{x}\int_{x}^{\infty} e^{-\xi}u(\xi)\,d\xi+f(x)$$

designate these integrals as $v(x)$ and $w(x)$, differentiate them, and so on. (Let the reader complete the reduction and find out why for $|\operatorname{Im} k_{1,2}|>1$ arbitrary constants do not enter into the solution of the integral equation while they occur in the solution of the differential equation.)

To solve equation (34) we sometimes use the **two-sided Laplace transformation**

$$F(p)=\int_{-\infty}^{\infty} e^{-px}f(x)\,dx$$

(cf. (III.1.1)) instead of the Fourier transformation. There is no essential difference between these transformations because we have

$$\hat{f}(k)=\frac{1}{2\pi}F(ik) \quad \text{that is} \quad \hat{f}(k)=\frac{1}{2\pi}F(p) \quad \text{where} \quad p=ik$$

Some equations are reduced to form (34) with the aid of an appropriately chosen substitution. For instance, the equation

$$u(x)=\lambda\int_0^{\infty} r\left(\frac{x}{\xi}\right)u(\xi)\frac{d\xi}{\xi}+f(x) \qquad (0<x<\infty) \tag{47}$$

is reduced to (34) by means of the substitution

$$x=e^{x_1},\ \xi=e^{\xi_1},\ u(e^{x_1})=u_1(x_1),\ r(e^{x_1})=r_1(x_1),\ f(e^{x_1})=f_1(x_1)$$

(check it up!). Since it is the substitution that reduces Mellin's transformation (Sec. III.3.3) to Fourier's transformation, we can directly apply the former to equation (47), which transforms it to equation (28). We shall not dwell on this question here.

In order to solve the equation

$$u(x)=\lambda\int_{-\infty}^{\infty} r(x+\xi)\,u(\xi)\,d\xi+f(x) \tag{48}$$

we replace ξ by $-\xi$ and put $u(-x)=v(x)$, $r(-x)=\rho(x)$. Then from (48) we obtain the system of equations (check it up!)

$$\left.\begin{aligned} u(x) &= \lambda \int_{-\infty}^{\infty} r(x-\xi)\, v(\xi)\, d\xi + f(x), \\ v(x) &= \lambda \int_{-\infty}^{\infty} \rho(x-\xi)\, u(\xi)\, d\xi + f(-x) \end{aligned}\right\}$$

which are of form (34). Fourier's transformation can be applied to this system both for absolutely integrable and for increasing functions, which we leave to the reader.

8. **Fredholm's One-Dimensional Equations with Difference Kernels on the Semi-Axis.** Let us consider the equation

$$u(x) = \lambda \int_{0}^{\infty} r(x-\xi)\, u(\xi)\, d\xi \qquad (0 \leqslant x < \infty) \tag{49}$$

It turns out that using some new considerations, namely, the so-called **Wiener-Hopf method**, we can apply Fourier's transformation to equation (49) as well.

First of all it should be noted that though the function $r(x)$ must be defined on the whole x-axis, the solution $u(x)$, according to the meaning of the problem, is constructed only for $x \geqslant 0$. However, if, after finding this solution, we put $u(x)$ equal to the right-hand side of (49) for $x < 0$, we obtain a function satisfying equation (49) on the whole x-axis. Therefore, we shall consider $u(x)$ as defined for all x. Using the notation of Sec. 7 we can write

$$u_+(x) + u_-(x) = \lambda \int_{-\infty}^{\infty} r(x-\xi)\, u_+(\xi)\, d\xi \qquad (-\infty < x < \infty) \tag{50}$$

As in Sec. 7, we shall assume that the function $r(x)$ satisfies an inequality of form (39) with exponents $\beta_\pm$ ($\beta_+ < \beta_-$) and that for the solution $u(x)$ the estimation $|u(x)| \leqslant Ae^{\gamma_+ x}$ $(0 \leqslant x < \infty)$ holds where $\beta_+ < \gamma_+ < \beta_-$. Then, by the lemma of Sec. 7, we obtain

$$|u(x)| \leqslant Be^{\beta_- x} \qquad (-\infty < x \leqslant 0)$$

Passing to Fourier's transforms in equation (50) we receive

$$\hat{u}_+(k) + \hat{u}_-(k) = 2\pi\lambda \hat{r}(k)\, \hat{u}_+(k), \text{ i.e. } (1 - 2\pi\lambda \hat{r}(k))\, \hat{u}_+(k) = -\hat{u}_-(k) \tag{51}$$

The domains in which the analyticity of the functions involved can be guaranteed are shown in Fig. 116. Each of the functions

$\hat{u}_+(k)$ and $\hat{u}_-(k)$ tends to zero for $k \to \infty$ within its domain of analyticity (why?).

The main idea of the Wiener-Hopf method is based on the **factorization**: the coefficient $1-2\pi\lambda\hat{r}(k)$ is represented as a product of factors which possess some preassigned properties. If such factorization can be realized, two unknown functions $\hat{u}_+$ and $\hat{u}_-$

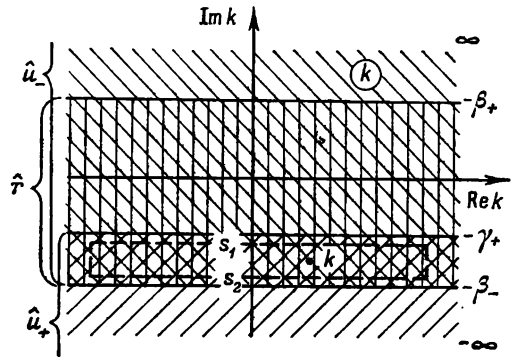


Fig. 116

can be found from one equation (51). Thus, let us suppose that it is possible to write the coefficient $1-2\pi\lambda\hat{r}(k)$ in the form

$$1-2\pi\lambda\hat{r}(k)=\frac{\varphi_1(k)}{\varphi_2(k)} \qquad (-\beta_- < \operatorname{Im} k < -\gamma_+)$$

where the function $\varphi_1(k)$ is analytic in the half-plane $\operatorname{Im} k < -\gamma_+$ and $\varphi_2(k)$ in the half-plane $\operatorname{Im} k > -\beta_-$, both functions increasing not faster than k^m $(m = 1, 2, 3, \ldots)$ as $k \to \infty$. Rewriting equation (51) in the form

$$\varphi_1(k)\,\hat{u}_+(k) = -\varphi_2(k)\,\hat{u}_-(k) \tag{52}$$

we see that each side of this equation is the analytic continuation of the other (see Sec. II.2.7), and therefore they form a (complete) analytic function $P(k)$ defined in the entire complex k-plane. This function is of order $l < m$ relative to k for $k \to \infty$ and thus is a polynomial of degree not higher than m (Sec. II.3.6). From (52), using the inversion formula, we obtain the solution of equation (49):

$$u(x) = \int_{-\infty+is}^{\infty+is} \frac{P(k)}{\varphi_1(k)} e^{ikx}\,dk \qquad (0 < x < \infty,\ s < -\gamma_+)$$

The factorization is particularly simple if the function $1-2\pi\lambda\hat{r}(k)$ is rational. Then it is possible to factorize its nume-

rator and denominator into factors of the form $k-a$ and to construct the function φ_1 as the product of all factors of the numerator involving $\operatorname{Im} a > -\beta_-$ and all factors of the denominator with $\operatorname{Im} a \geqslant -\gamma_+$. In this case equation (49) can be reduced, by analogy with Sec. 7, to an ordinary linear differential equation with constant coefficients.

In the general case the factorization can be performed in the following way. First of all, if the function $1-2\pi\lambda\hat{r}(k)$ has no zeros for

$$-\beta_- < \operatorname{Im} k < -\gamma_+ \tag{53}$$

it is written in the form $v(k) = e^{w(k)}$ $(w(k) = \ln(1-2\pi\lambda\hat{r}(k))$ after which $w(k)$ is represented as a difference of functions analytic in the corresponding half-planes. To this end, we can apply Cauchy's integral formula (II.3.12) to the contour of integration shown in Fig. 116 in dotted line. If the lateral sides of the contour are made to recede to infinity (show that the integrals over these sides then tend to zero) we obtain, in the limit, the expression

$$w(k) = \frac{1}{2\pi i}\int\limits_{-\infty+is_1}^{\infty+is_1} \frac{w(\varkappa)}{k-\varkappa}\,d\varkappa - \frac{1}{2\pi i}\int\limits_{-\infty+is_2}^{\infty+is_2} \frac{w(\varkappa)}{k-\varkappa}\,d\varkappa \tag{54}$$

This is the required representation. If the function $1-2\pi\lambda\hat{r}(k)$ has zeros a_j of multiplicities n_j in the strip (53) it can be written in the form $\frac{Q(k)}{R(k)}v(k)$ where $Q(k) = \prod_j (k-a_j)^{n_j}$ and $R(k)$ is an arbitrary polynomial equivalent to $Q(k)$ for $k \to \infty$ and having no zeros in the strip (53). Then we must write $v(k) = e^{w(k)}$, etc., and transform the ratio $\frac{Q(k)}{R(k)}$ as was done above for $\frac{\varphi_1(k)}{\varphi_2(k)}$.

For nonhomogeneous equation (49) we can in an arbitrary manner extend the function $f(x)$ to the interval $-\infty < x < 0$ and then, by analogy with the above, obtain, instead of (52), the equation

$$\varphi_1(k)\,\hat{u}_+(k) - \varphi_2(k)\,\hat{f}_+(k) = [-\hat{u}_-(k) + \hat{f}_-(k)]\,\varphi_2(k) \tag{55}$$

Now, applying the procedure described above we write the subtrahend in the form of the difference

$$\varphi_2(k)\,\hat{f}_+(k) = \psi_1(k) - \psi_2(k) \tag{56}$$

where the function $\psi_1(k)$ is analytic for $\operatorname{Im} k < -\gamma_+$ and $\psi_2(k)$ is analytic for $\operatorname{Im} k > -\beta_-$, both functions increasing for $k \to \infty$ not faster than a power function. From (55) we obtain

$$\varphi_1(k)\,\hat{u}_+(k) - \psi_1(k) = [-\hat{u}_-(k) + \hat{f}_-(k)]\,\varphi_2(k) - \psi_2(k)$$

and then reason in the same way as in the case of equation (52).

The Wiener-Hopf method can also be used for solving integral equations of the first kind having an analogous structure.

As an example, let us consider **Milne's equation** (E. Milne, (1896-1950), an English physicist):

$$\rho(x)=\int_0^\infty r(x-\xi)\rho(\xi)\,d\xi \text{ where } r(x)=\frac{1}{2}\int_0^1 e^{\frac{-|x|}{s}}\frac{ds}{s}$$

This equation is encountered in physics; its solution describes the intensity of radiation in the half-space $0\leqslant x<\infty$, $-\infty<y$, $z<\infty$ propagating in an isotropic homogeneous substance which does not absorb the radiation but scatters it, x being the dimensionless distance (related to the free-path length).

The computation yields (check it up!)

$$\hat{f}(k)=\frac{1}{2\pi}\int_{-\infty}^{\infty}dx\,e^{-ikx}\frac{1}{2}\int_0^1 e^{\frac{-|x|}{s}}\frac{ds}{s}=$$

$$=\frac{1}{4\pi}\int_0^1\frac{ds}{s}\left(\int_{-\infty}^0\exp\left(-\frac{|x|}{s}-ikx\right)dx+\int_0^\infty\exp\left(-\frac{|x|}{s}-ikx\right)dx\right)=$$

$$=\frac{1}{4\pi}\int_0^1\left(\frac{1}{1-iks}+\frac{1}{1+iks}\right)ds=\frac{1}{4i\pi k}\ln\frac{1+ik}{1-ik}$$

Therefore, it is necessary to factor the function

$$1-2\pi\hat{r}(k)=1-\frac{1}{2ik}\ln\frac{1+ik}{1-ik} \tag{57}$$

in the strip

$$-1<\operatorname{Im}k<1 \tag{58}$$

Passing in (57) to the difference of the logarithms and using Taylor's series we see that for $k=0$ there is a zero of order 2. It can be proved that function (57) has no other zeros in the strip (58). Therefore, we can rewrite (57) as

$$1-2\pi\hat{r}(k)=\frac{k^2}{k^2+1}\left[\frac{k^2+1}{k^2}\left(1-\frac{1}{2ik}\ln\frac{1+ik}{1-ik}\right)\right]$$

whence

$$w(k)=\ln\left[\frac{k^2+1}{k^2}\left(1-\frac{1}{2ik}\ln\frac{1+ik}{1-ik}\right)\right]$$

In formula (54) we can put $s_1=0$, which results in

$$\varphi_1(k)=\frac{k^2}{k-i}\exp\frac{1}{2\pi i}\int_{-\infty}^{\infty}\frac{1}{k-\varkappa}\ln\left[\frac{\varkappa^2+1}{\varkappa^2}\left(1-\frac{\arctan\varkappa}{\varkappa}\right)\right]d\varkappa \quad (\operatorname{Im}k<0)$$

Since for $k \to -i\infty$ this function is equivalent to k, and $u_+(k) \to 0$, we have $P(k) = o(k)$, and hence the polynomial $P(k)$ is equal to a constant. It follows that

$$\hat{u}_+(k) = C\frac{k-i}{k^2}\exp\frac{-1}{2\pi i}\int_{-\infty}^{\infty}\frac{1}{k-\varkappa}\ln\left[\frac{\varkappa^2+1}{\varkappa^2}\left(1-\frac{\arctan\varkappa}{\varkappa}\right)\right]d\varkappa \quad (\operatorname{Im} k < 0)$$

The Wiener-Hopf method can be applied to some equations similar to (49). For example, let us consider the equation

$$u(x) = \begin{cases} \int_{-\infty}^{\infty} r_1(x-\xi)\,u(\xi)\,d\xi + f(x) & (-\infty < x < 0) \\ \int_{-\infty}^{\infty} r_2(x-\xi)\,u(\xi)\,d\xi + f(x) & (0 < x < \infty) \end{cases}$$

Putting

$$v(x) = \begin{cases} u(x) - \int_{-\infty}^{\infty} r_2(x-\xi)\,u(\xi)\,d\xi & (-\infty < x < 0) \\ u(x) - \int_{-\infty}^{\infty} r_1(x-\xi)\,u(\xi)\,d\xi & (0 < x < \infty) \end{cases}$$

we can rewrite the given equation as

$$u(x) = \int_{-\infty}^{\infty} r_1(x-\xi)\,u(\xi)\,d\xi + v_+(x) + f_-(x) =$$

$$= \int_{-\infty}^{\infty} r_2(x-\xi)\,u(\xi)\,d\xi + v_-(x) + f_+(x) \qquad (-\infty < x < \infty)$$

(check it up!) which is a system of two equations. Under certain requirements (let the reader establish the corresponding conditions on the orders of growth of the functions involved), we can pass to Fourier's transforms, which leads to

$$\hat{u}(k) = 2\pi\hat{r}_1(k)\,\hat{u}(k) + \hat{v}_+(k) + \hat{f}_-(k) = 2\pi\hat{r}_2(k)\,\hat{u}(k) + \hat{v}_-(k) + \hat{f}_+(k) \tag{59}$$

Eliminating $\hat{u}(k)$ we obtain the relation

$$\frac{1-2\pi\hat{r}_2(k)}{1-2\pi\hat{r}_1(k)}\left[\hat{v}_+(k) + \hat{f}_-(k)\right] = \hat{v}_-(k) + \hat{f}_+(k)$$

which must be fulfilled in the corresponding strip. Now it only remains to factorize the quotient $\frac{1-2\pi\hat{r}_2(k)}{1-2\pi\hat{r}_1(k)}$ and, if the equations are nonhomogeneous, to write down representations of type (56),

after which considerations similar to those concerning equation (52) enable us to compute $\hat{v}_+(k)$ and then, with the aid of (59), determine the function $\hat{u}(k)$.

§ 5. SINGULAR INTEGRAL EQUATIONS

The theory of singular integral equations in which unknown functions enter into singular integrals is more sophisticated than that of Fredholm's equations. Here we shall limit ourselves to the simplest case of one-dimensional scalar equations. For more detail we refer the reader to [98], [121], [130], and [139] and also to special monographs [101], [106], [135].

1. Singular Integrals. Here we shall deal with integrals of the form

$$\int_a^b \frac{\varphi(\xi)}{\xi - x}\, d\xi \tag{1}$$

and also with some other similar integrals. We shall begin with the case when a and b are finite. As is known (e.g. see [107], § XIV.4) in the general case integral (1) is divergent in the vicinity of the point $\xi = x$. If $a < x < b$, its principal value is defined by the formula

$$\text{v.p.} \int_a^b \frac{\varphi(\xi)}{\xi - x}\, d\xi = \lim_{\varepsilon \to +0} \left(\int_a^{x-\varepsilon} + \int_{x+\varepsilon}^b \right) \frac{\varphi(\xi)}{\xi - x}\, d\xi \tag{2}$$

Using the formula

$$\text{v.p.} \int_a^b \frac{1}{\xi - x}\, d\xi = \ln \left| \frac{b-x}{a-x} \right|$$

(prove it!) and transforming integral (1) we obtain

$$\int_a^b \frac{\varphi(\xi)}{\xi - x}\, d\xi = \int_a^b \frac{\varphi(\xi) - \varphi(x)}{\xi - x}\, d\xi + \int_a^b \frac{\varphi(x)}{\xi - x}\, d\xi$$

which yields

$$\text{v.p.} \int_a^b \frac{\varphi(\xi)}{\xi - x}\, d\xi = \varphi(x) \ln \frac{b-x}{x-a} + \int_a^b \frac{\varphi(\xi) - \varphi(x)}{\xi - x}\, d\xi \tag{3}$$

For a continuous function $\varphi(\xi)$ differentiable at the point $\xi = x$ the last integral is proper (why?) and, consequently, it takes on a finite value. By the way, it can be verified that for continuous nondifferentiable functions this integral, even if it is improper, is convergent (except some artificial examples which are of no prac-

tical interest) and depends continuously on x. But if the function $\varphi(\xi)$ is discontinuous for $\xi = x$, the simplest examples show that for $\xi = x$ integral (3) may be infinite even when the function $\varphi(\xi)$ is finite. Furthermore, the first term on the right-hand side of (3) indicates that the left-hand side tends to infinity with the rate of a logarithmic function for $x = a + 0$ if $\varphi(a) \neq 0$ and $\varphi(a) \neq \pm\infty$; the same applies to $x = b - 0$.

These properties are readily demonstrated (let the reader perform the calculations) by the example of the function

$$\varphi(\xi) = \begin{cases} 0 \text{ for } -1 \leqslant \xi < 0 \\ 1 \text{ for } \quad 0 < \xi \leqslant 1 \end{cases}$$

for which

$$\text{v.p.} \int_{-1}^{1} \frac{\varphi(\xi)}{\xi - x} d\xi = \ln \frac{1-x}{|x|} \qquad (-1 \leqslant x \leqslant 1)$$

Here is another useful formula:

$$\text{v.p.} \int_{-1}^{1} \left(\frac{1-\xi}{1+\xi}\right)^{\alpha} \frac{d\xi}{\xi - x} = \pi \cot \alpha\pi \left(\frac{1-x}{1+x}\right)^{\alpha} - \frac{\pi}{\sin \pi\alpha} \qquad (-1 < \alpha < 1, \alpha \neq 0) \tag{4}$$

(Let the reader derive it from (II.3.28) taking into account the remark in Sec. II.3.4 according to which for the singular points lying on the contour of integration the residues are taken with the factor $\frac{1}{2}$ (there are two such points in this case because the point $z = x$ can be approached both from above and from below).) From formula (4) it is readily seen that if the function $\varphi(\xi)$ is of the order of $(\xi - a)^{-\alpha}$ $((\xi - b)^{-\alpha})$ in the vicinity of the point $\xi = a$ $(\xi = b)$ where $0 < \alpha < 1$ integral (1) as function of x is of the same order.

The integral

$$\int \varphi(\xi) \cot \frac{\xi - x}{2} d\xi \tag{5}$$

where $\varphi(\xi)$ is a function with period 2π and the integration extends over any interval of length 2π, possesses analogous properties since the difference $\cot \frac{\xi - x}{2} - \frac{2}{\xi - x}$ remains finite for $\xi = x$.

In the case $a = -\infty$ or $b = \infty$ we can split the interval of integration into a finite part containing the point $\xi = x$ and the remaining infinite part; if integral (1) over the infinite part is absolutely convergent, the properties of the original integral are completely determined by its singular part.

We shall also use complex integrals

$$\frac{1}{\pi i}\int_{(L)}\frac{\varphi(\zeta)}{\zeta-z}d\zeta \tag{6}$$

similar to (1) where (L) is a finite oriented closed or nonclosed contour in the complex plane, and $\varphi(\xi)$ is a function defined on this contour. If $z\in(L)$ integral (6) has the same properties as (1) but, besides, it can turn into infinity at the corner points of the contour provided there are such (let the reader construct simple illustrative examples demonstrating these properties). If $z\overline{\in}(L)$ and the function φ is continuous or at least absolutely integrable, integral (6) becomes regular. Since integral (6) can be differentiated with respect to the parameter z we see that $\Phi(z)=\frac{1}{\pi i}\int_{(L)}\frac{\varphi(\zeta)}{\zeta-z}d\zeta$ is a one-valued analytic function of z everywhere except the curve (L) which, as is shown below, is a line of discontinuity of the function.

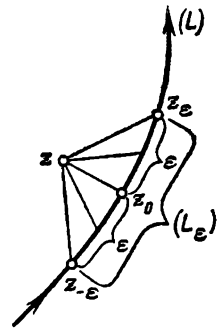


Fig. 117

Indeed, let the moving point $z\overline{\in}(L)$ be in the vicinity of a point $z_0\in(L)$. Choose some $\varepsilon>0$ and denote by (L_ε) the arc shown in Fig. 117, and by (L'_ε) the remaining part of the line (L). Then

$$\Phi(z)=\frac{1}{\pi i}\int_{(L')}\frac{\varphi(\zeta)}{\zeta-z}d\zeta+\frac{1}{\pi i}\int_{(L_\varepsilon)}\frac{\varphi(\zeta)}{\zeta-z}d\zeta$$

For all points z lying sufficiently close to z_0 the first summand is approximately equal to $\frac{1}{\pi i}$ v.p. $\int_{(L)}\frac{\varphi(\zeta)}{\zeta-z_0}d\zeta$. In the second summand we can approximately put $\varphi(\zeta)=\varphi(z_0)$; then the constant $\varphi(z_0)$ can be taken outside the integral sign, and the remaining integral is equal to the increment of the expression $\ln(\zeta-z)$ gained as ζ traces (L_ε) from $z_{-\varepsilon}$ to z_ε. But this increment is approximately equal to πi if z is, as in Fig. 117, to the left of z_0 when the contour (L) is traced according to its orientation and to $-\pi i$ if z is to the right of z_0. Denoting by Φ^l and Φ^r the limiting values $\Phi(z)$ attained when $z\overline{\in}(L)$ approaches $z_0\in(L)$ from the left and from the right, respectively, and passing to the limit for $\varepsilon\to 0$ we receive

$$\begin{aligned}\Phi^l(z_0)&=\frac{1}{\pi i}\text{ v. p.}\int_{(L)}\frac{\varphi(\zeta)}{\zeta-z_0}d\zeta+\varphi(z_0),\\ \Phi^r(z_0)&=\frac{1}{\pi i}\text{ v. p.}\int_{(L)}\frac{\varphi(\zeta)}{\zeta-z_0}d\zeta-\varphi(z_0)\end{aligned} \tag{7}$$

which means that (L) is in fact a line of discontinuity of $\Phi(z)$.

When dealing with singular integrals we must take into account their specific nature. For instance, if it is necessary to differentiate integral (1) with respect to x, it cannot be done by Leibniz' rule because then we would obtain an integral without a principal value. Therefore, using definition (2), we first integrate by parts and then perform the differentiation, which shows that the derivative is equal to $\text{v.p.} \int_a^b \frac{\varphi'(\xi)\,d\xi}{\xi - x} + \frac{\varphi(b)}{x-b} - \frac{\varphi(a)}{x-a}$. Here is another example: it can be proved that if a multiple integral involves a singular integration and a nonsingular one their order can be reversed in the ordinary way but if there are two singular integrations we must use the formula

$$\text{v.p.} \int_{(L)} dz \left(\text{v.p.} \int_{(L)} \frac{F(z, \zeta)}{(z-c)(\zeta - z)}\, d\zeta \right) =$$

$$= \text{v.p.} \int_{(L)} d\zeta \left(\text{v.p.} \int_{(L)} \frac{F(z, \zeta)}{(z-c)(\zeta - z)}\, dz \right) - \pi^2 F(c, c) \qquad (c \in (L))$$

2. Inversion Formulas. One of the most important methods of studying singular integral equations is based on the application of inversion formulas for integral transformations. Here we remind the reader of the inversion formulas derived in Secs. II.3.5 and III.3.3.

The transformation with the **Cauchy kernel** $(\xi - x)^{-1}$: if

$$F(x) = \frac{1}{\pi} \text{v.p.} \int_{-\infty}^{\infty} \frac{f(\xi)}{\xi - x}\, d\xi$$

then

$$f(x) = -\frac{1}{\pi} \text{v.p.} \int_{-\infty}^{\infty} \frac{F(\xi)}{\xi - x}\, d\xi$$

The transformation with the **Hilbert kernel**: if

$$F(x) = \frac{1}{2\pi} \text{v.p.} \int f(\xi) \cot \frac{\xi - x}{2}\, d\xi \qquad (8)$$

then

$$f(x) = -\frac{1}{2\pi} \text{v.p.} \int F(\xi) \cot \frac{\xi - x}{2}\, d\xi + \frac{1}{2\pi} \int f(\xi)\, d\xi$$

(when deriving the latter formula, one should take into account both (II.3.38) and (II.3.37)). In this transformation f and F are periodic functions with period 2π and the integrals are taken over arbitrary fixed intervals of length 2π.

Let us derive the inversion formula for integral (6) applicable to the case of a closed contour (L). We thus consider the function

$$F(z)=\frac{1}{\pi i}\ \text{v.p.}\oint\limits_{(L)}\frac{f(\zeta)}{\zeta-z}d\zeta \qquad (z\in(L)) \tag{9}$$

To obtain the formula we put

$$\varphi(z_1)=\frac{1}{\pi i}\oint\limits_{(L)}\frac{f(\zeta)}{\zeta-z_1}d\zeta,\quad \Phi(z_1)=\frac{1}{\pi i}\oint\limits_{(L)}\frac{F(\zeta)}{\zeta-z_1}d\zeta\ \ (z_1\ \overline{\in}(L)) \tag{10}$$

We shall consider (L) to be oriented counterclockwise (for the ultimate result this assumption is inessential). Let z_1 vary in an arbitrary manner inside (L). As was shown, φ and Φ are analytic functions taking on, respectively, the limiting values

$$\varphi^1(z)=\frac{1}{\pi i}\ \text{v.p.}\oint\limits_{(L)}\frac{f(\zeta)}{\zeta-z}d\zeta+f(z)=F(z)+f(z) \tag{11}$$

and

$$\Phi^1(z)=\frac{1}{\pi i}\ \text{v.p.}\oint\limits_{(L)}\frac{F(\zeta)}{\zeta-z}d\zeta+F(z) \tag{12}$$

for $z\in(L)$. Expressing $F(z)$ from (11) and substituting it into (10) we obtain

$$\Phi(z_1)=\frac{1}{\pi i}\oint\limits_{(L)}\frac{\varphi^1(\zeta)}{\zeta-z_1}d\zeta-\frac{1}{\pi i}\oint\limits_{(L)}\frac{f(\zeta)}{\zeta-z_1}d\zeta=\frac{1}{\pi i}\oint\limits_{(L)}\frac{\varphi^1(\xi)}{\zeta-z_1}d\zeta-\varphi(z_1)$$

However, by Cauchy's integral formula (II.3.12), the integral on the right-hand side is equal to $2\varphi(z_1)$. Hence, $\Phi(z_1)\equiv\varphi(z_1)$ and $\Phi^1(z)=\varphi^1(z)$, and from (11) and (12) we receive

$$f(z)=\frac{1}{\pi i}\ \text{v.p.}\oint\limits_{(L)}\frac{F(\zeta)}{\zeta-z}d\zeta \tag{13}$$

This is the desired inversion formula for transformation (9). It means that this transformation is the inverse of itself. We have thus proved the so-called **Poincaré-Bertrand theorem** (J.H. Poincaré (1854-1922) and J.L.F. Bertrand (1822-1900), French mathematicians).

A more thorough analysis shows that if the original function is square-summable, that is belongs to L_2, its image also belongs to L_2.

Here we give without proof one more inversion formula: if

$$F(x)=\frac{1}{\pi}\ \text{v.p.}\int\limits_{-1}^{1}\frac{f(\xi)}{\xi-x}d\xi \qquad (-1<x<1) \tag{14}$$

then

$$f(x) = -\frac{1}{\pi}\ \text{v.p.}\int_{-1}^{1}\sqrt{\frac{1-\xi^2}{1-x^2}}\,\frac{F(\xi)}{\xi-x}\,d\xi+\frac{C}{\sqrt{1-x^2}} \tag{15}$$

where C is a constant (see Sec. 6). We shall only check that for any C the substitution of the last term into (14) results in zero, that is

$$\text{v.p.}\int_{-1}^{1}\frac{d\xi}{(\xi-x)\sqrt{1-\xi^2}}=0\quad(-1<x<1)$$

This can be proved by means of formula (2) as the indefinite integral is readily computed, but it is easier to make the substitution $\xi=\cos\varphi$, $x=\cos 0$, after which we take advantage of the formulas derived at the end of Sec. II.3.5, which yields the required result. (It should be noted that in the integrals of form (1) the change of variables is carried out according to the ordinary formulas.)

If we regard formula (9) as a definition of a singular integral operator $\mathbf{F}=\mathbf{S}f$ in the space $L_2(\Gamma)$ we can prove that it is bounded. At the same time, from the inversion formula we obtain the relation $S^2=I$, and hence it follows that the operator is not completely continuous (Sec. 2.11). Indeed, the product of completely continuous operators is always completely continuous (why?), whereas the identity operator in an infinite-dimensional space is not completely continuous. Here lies a substantial distinction between singular and Fredholm's operators, which, as will be seen, essentially affects the formulations of theorems on solvability of singular integral equations.

3. Direct Application of Inversion Formulas. Let us consider the equation

$$\alpha u(z)=\frac{\beta}{\pi i}\ \text{v.p.}\oint_{(L)}\frac{u(\zeta)}{\zeta-z}\,d\zeta+f(z) \tag{16}$$

where α and β are constants and $\alpha^2\neq\beta^2$. Applying transformation (9) to both sides and taking into account inversion formula (13) we obtain

$$\frac{\alpha}{\pi i}\ \text{v.p.}\oint_{(L)}\frac{u(\zeta)}{\zeta-z}\,d\zeta=\beta u(z)+F(z) \tag{17}$$

Eliminating the integral over (L) from relations (16) and (17) we arrive at the formula

$$u(z)=\frac{1}{\alpha^2-\beta^2}\left[\alpha f(z)+\beta F(z)\right] \tag{18}$$

where the function $F(z)$ is determined by formula (9). We see

that for $\alpha \to 0$ the solution remains finite while for $\alpha \to \pm\beta$ it can have a singularity.

Applying the same transformation to the more general equation

$$\alpha u(z) = \frac{\beta}{\pi i} \text{v.p.} \oint_{(L)} \frac{u(\zeta)}{\zeta - z} d\zeta + \oint_{(L)} K(z, \zeta) u(\zeta) d\zeta + f(z) \qquad (19)$$

with a Fredholm kernel K we obtain a relation of form (18) in which f should be replaced by the sum of the last two terms of equation (19) and the function F should be appropriately changed. Reversing the order of integration in the double integral we obtain the resultant relation

$$u(z) = \frac{1}{\alpha^2 - \beta^2} \left[\alpha f(z) + \beta F(z) + \oint_{(L)} K_1(z, \zeta) u(\zeta) d\zeta \right] \qquad (20)$$

where

$$K_1(z, \zeta) = \alpha K(z, \zeta) + \frac{\beta}{\pi i} \text{v.p.} \oint_{(L)} \frac{K(w, \zeta)}{w - z} dw \qquad (21)$$

(let the reader carry out the calculations taking into account that, as was mentioned in Sec. 1, it is allowable to reverse the order of singular and ordinary integrations). It can be proved that the kernel $K_1(z, \zeta)$ defined by formula (21) is nonsingular and specifies a Fredholm-type operator. We have thus reduced the original singular equation (19) to Fredholm's equation (20), that is performed the so-called **regularization** of equation (20). Since all the results of § 2 are applicable to equation (20), we conclude, in particular, that three Fredholm's theorems hold for the solutions of equation (19).

In connection with these theorems we note that the adjoint kernel of $K(z, \zeta)$ is written in the form $[K(\zeta, z)]^* \frac{dz^*}{|dz|} \frac{|d\zeta|}{d\zeta}$. Prove it on the basis of the definition of the adjoint operator (see (2.66)) and of the definition of the scalar product in $L_2(L)$ which is expressed by the formula

$$(\mathbf{f}, \varphi)_{L_2(L)} = \int_{(L)} f(z) \varphi^*(z) |dz|$$

The equation

$$\alpha u(x) = \frac{\beta}{\pi} \text{v.p.} \int_{-\infty}^{\infty} \frac{u(\xi)}{\xi - x} d\xi + \int_{-\infty}^{\infty} K(x, \xi) u(\xi) d\xi + f(x)$$

$$(\alpha^2 + \beta^2 \neq 0)$$

is investigated similarly.

If transformation (15) is applied to the equation

$$\alpha u(x) = \frac{\beta}{2\pi} \text{ v.p.} \int u(\xi) \cot \frac{\xi - x}{2} d\xi + \int K(x, \xi) u(\xi) d\xi + f(x)$$

$$(\alpha^2 + \beta^2 \neq 0)$$

(where all the functions involved are periodic with period 2π) the first term on the right-hand side yields the additional summand $-\frac{\beta}{2\pi} \int u(\xi) d\xi$ which should be included in the expression containing the transformed kernel K_1.

To solve the equation

$$\alpha u(x) = \frac{\beta}{\pi} \text{ v. p.} \int_{-1}^{1} \frac{u(\xi)}{\xi - x} d\xi + \int_{-1}^{1} K(x, \xi) u(\xi) d\xi + f(x) \qquad (22)$$

$$(\alpha^2 + \beta^2 \neq 0)$$

we replace ξ by s and x by ξ, multiply both members by $\frac{1}{\pi} \sqrt{\frac{1-\xi^2}{1-x^2}} \frac{1}{\xi - x}$ and integrate with respect to ξ from -1 to 1. Then, by virtue of inversion formula (15), we obtain

$$\frac{\alpha}{\pi} \text{ v.p.} \int_{-1}^{1} \sqrt{\frac{1-\xi^2}{1-x^2}} \frac{u(\xi)}{\xi - x} d\xi = -\beta u(x) + \frac{C}{\sqrt{1-x^2}} + \ldots$$

Now we represent the left member as the sum

$$\frac{\alpha}{\pi} \text{ v.p.} \int_{-1}^{1} \frac{u(\xi)}{\xi - x} d\xi + \frac{\alpha}{\pi} \int_{-1}^{1} \left(\sqrt{\frac{1-\xi^2}{1-x^2}} - 1 \right) \frac{u(\xi)}{\xi - x} d\xi$$

where the first integral is singular and the second regular (check it up!) and eliminate the singular integral from the equation thus obtained and from the original equation. Hence, equation (22) can also be regularized.

4. Reduction to a Boundary-Value Problem. Example. Let us consider the equation

$$\alpha(z) u(z) = \frac{\beta(z)}{\pi i} \text{ v.p.} \int_{(L)} \frac{u(\zeta)}{\zeta - z} d\zeta + f(z) \qquad (23)$$

$$(z \in (L), \ \alpha^2(z) - \beta^2(z) \neq 0)$$

generalizing (16) where $\alpha(z)$ and $\beta(z)$ are continuous coefficients and (L) is a closed or nonclosed finite contour. One of the main methods of studying this equation which in many cases makes it possible to obtain its solution consists in passing to an equivalent boundary-value problem of the theory of analytic functions. This method is thoroughly studied in [43].

Let $u(z)\,(z \in (L))$ be the sought-for solution. For our further aims we introduce the function

$$U(z_1) = \frac{1}{2\pi i}\int_{(L)} \frac{u(\zeta)}{\zeta - z_1}\,d\zeta \qquad (z_1 \overline{\in} (L))$$

As was shown in Sec. 1, this is a one-valued analytic function in the entire z_1-plane except the contour (L); it is also clear that $U(\infty) = 0$. For the points of the contour (L) we have the relations

$$\left.\begin{aligned} U^{1}(z) &= \frac{1}{2\pi i}\,\text{v.p}\int_{(L)} \frac{u(\zeta)}{\zeta - z}\,d\zeta + \frac{u(z)}{2} \\ U^{r}(z) &= \frac{1}{2\pi i}\,\text{v.p.}\int_{(L)} \frac{u(\zeta)}{\zeta - z}\,d\zeta - \frac{u(z)}{2} \end{aligned}\right\} \qquad (24)$$

(see (7)). Eliminating $u(z)$ and the integral from these two equalities and equation (23) we obtain the relation (check it up!)

$$U^{1}(z) = \frac{\alpha(z)+\beta(z)}{\alpha(z)-\beta(z)}\,U^{r}(z) + \frac{f(z)}{\alpha(z)-\beta(z)} \qquad (z \in (L)) \qquad (25)$$

Thus we arrive at the boundary-value problem known as the **Riemann-Hilbert problem**: it is required to find a function $U(z_1)$ which is one-valued and analytic in the entire z_1-plane except the line of discontinuity (L), equal to zero for $z_1 = \infty$ and satisfies boundary condition (25) on (L). We can prove that, conversely, if $U(z_1)$ is a solution of this problem the function $U(z) = U^{1}(z) - U^{r}(z)$ (see (24)) is a solution of equation (23); thus, this equation is completely equivalent to the Riemann-Hilbert boundary-value problem.

Let us first suppose that (L) is a closed positively oriented contour. Then it is more natural to regard the function $U(z_1)$ as an aggregate of two functions $U^{1}(z_1)$ and $U^{r}(z_1)$ defined, respectively, inside and outside (L), and connected on (L) by relation (25). Specific features of the Riemann-Hilbert problem can be demonstrated by a simple example in which (L) is the circle $|z| = 1$ and condition (25) has the form

$$U^{1}(z) = z^{m}U^{r}(z) + f_1(z) \qquad (|z| = 1,\ m \text{ is an integer}) \qquad (26)$$

The function $U^{1}(z_1)$ which is analytic in the circle $|z_1| < 1$ can be expanded into Taylor's series and the function $U^{r}(z_1)$ which is analytic for $|z_1| > 1$ can be expanded into Laurent's series which, by condition $U^{r}(\infty) = 0$, only involves negative powers of z_1. Besides, let us expand $f_1(e^{i\varphi})$ into complex Fourier's series.

We thus obtain

$$U^{1}(z_1)=\sum_{k=0}^{\infty}a_k z_1^k, \quad U^{r}(z_1)=\sum_{k=1}^{\infty}b_k z_1^{-k}$$

$$f_1(e^{i\varphi})=\sum_{k=-\infty}^{\infty}c_k e^{ik\varphi}$$

where the coefficients c_k are given and a_k and b_k are sought for. Condition (26) yields

$$\sum_{k=0}^{\infty}a_k e^{ik\varphi}-\sum_{k=1}^{\infty}b_k e^{i(m-k)\varphi}=\sum_{k=-\infty}^{\infty}c_k e^{ik\varphi} \tag{27}$$

Now, depending on the value of the number m (referred to as the **index** of the problem), we arrive at the following possible cases:

If $m=0$ the problem has a single solution:

$$a_0=c_0, \quad a_k=c_k, \quad b_k=-c_{-k} \qquad (k=1,\ 2,\ 3,\ \ldots)$$

If $m>0$ the problem has an m-parameter family of solutions. Indeed, from (27) we receive

$$a_k=c_k \quad (k=m,\ m+1,\ m+2,\ \ldots), \quad b_k=-c_{m-k}$$
$$(k=m+1,\ m+2,\ \ldots)$$

and

$$a_0-b_m=c_0, \quad a_1-b_{m-1}=c_1, \quad \ldots, \quad a_{m-1}-b_1=c_{m-1}$$

The latter relations show that $a_0,\ a_1,\ \ldots,\ a_{m-1}$ can be set arbitrarily, and then $b_1,\ b_2,\ \ldots,\ b_m$ can be determined uniquely. In particular, the corresponding homogeneous problem (for $f_1(z)\equiv 0$) has m linearly independent solutions

$$U^{1}(z_1)=z_1^k, \quad U^{r}(z_1)=z_1^{k-m} \qquad (k=0,\ 1,\ \ldots,\ m-1)$$

If $m<0$ the left-hand side of (27) does not contain terms of the form $e^{-i\varphi},\ \ldots,\ e^{-im\varphi}$. Hence, *for the problem to be solvable in the case $m<0$ it is necessary and sufficient that $c_{-1}=c_{-2}=\ldots=c_{-m}=0$,* that is *the function $f_1(z)$ should satisfy the $|m|$ orthogonality relations*

$$\int_0^{2\pi} f(e^{i\varphi})\,e^{ik\varphi}\,d\varphi=0 \qquad (k=1,\ 2,\ \ldots,\ |m|)$$

If these relations are fulfilled the problem has exactly one solution

$$a_0=c_0, \quad a_k=c_k, \quad b_k=-c_{m-k} \qquad (k=1,\ 2,\ 3,\ \ldots)$$

5. The Case of an Arbitrary Closed Contour. Let us come back to the general problem (25) with an arbitrary closed contour (L). It turns out that its solution has the same properties as in the

case of problem (26). The index m of this problem is equal to the number of circuits the point $\frac{\alpha(z)+\beta(z)}{\alpha(z)-\beta(z)}$ makes about the origin when the point z traverses the contour (L).

For simplicity's sake, we shall suppose that the point $z=0$ is inside (L) (let the reader show that the general case can be reduced to the latter). Let us put

$$\mu(z)=\operatorname{Ln}\frac{\alpha(z)+\beta(z)}{\alpha(z)-\beta(z)}-m\operatorname{Ln}z \qquad (z\in(L)) \tag{28}$$

where the value of the logarithmic function is arbitrarily chosen at a point $z\in(L)$ and then continuously extended along (L). After the point z describes the contour (L) the minuend and the subtrahend in (28) gain the same increment $2\pi i m$ (why?), which means that the function $\mu(z)$ returns to its original value, and thus it is continuous and single-valued on (L). Let us introduce the functions

$$\psi(z_1)=\frac{1}{2\pi i}\oint_{(L)}\frac{\mu(\zeta)}{\zeta-z_1}d\zeta,\quad V(z_1)=U(z_1)e^{-\psi(z_1)}\quad (z_1\overline{\in}(L))$$

which are single-valued and analytic on the entire z_1-plane except the contour (L). Then, using the equalities

$$\psi^1(z)=\frac{1}{2\pi i}\,\text{v.p.}\oint_{(L)}\frac{\mu(\zeta)}{\zeta-z}d\zeta+\frac{1}{2}\mu(z)$$

$$\psi^r(z)=\frac{1}{2\pi i}\,\text{v.p.}\oint_{(L)}\frac{\mu(\zeta)}{\zeta-z}d\zeta-\frac{1}{2}\mu(z)$$

$$V^1(z)=U^1(z)e^{-\psi^1(z)},\quad V^r(z)=U^r(z)e^{-\psi^r(z)}\qquad (z\in(L))$$

we can transform condition (25) to the form (check it up!)

$$V^1(z)=z^mV^r(z)+f_1(z)\qquad (z\in(L)) \tag{29}$$

where

$$f_1(z)=\frac{f(z)}{\alpha(z)-\beta(z)}\exp\left[-\frac{1}{2\pi i}\,\text{v.p.}\oint_{(L)}\frac{\mu(\zeta)}{\zeta-z}d\zeta-\frac{1}{2}\mu(z)\right]$$

Thus, for an arbitrary contour we obtained a condition of form (26) for the function V. It is natural that the solvability of the problem essentially depends on its index m.

Let $m=0$. Then problem (29) is solvable and its solution is

$$V(z_1)=\frac{1}{2\pi i}\oint_{(L)}\frac{f_1(\zeta)}{\zeta-z_1}d\zeta \tag{30}$$

(check it up!). This solution is unique. Indeed, the difference $W(z_1)$ of two such solutions must satisfy the condition $W^1(z_1)=W^r(z_1)$ on (L) and, consequently, this function is analytic in the entire

z_1-plane including (L) (Sec. II.2.5); however, the function $W(z_1)$ is equal to zero for $z_1=\infty$ and hence it must be equal to zero identically (why?). Thus, in the case $m=0$ problem (25), and therefore equation (23), have a unique solution. Taking into account the equalities

$$V^{\mathrm{I}}(z)=\frac{1}{2\pi i}\,\mathrm{v.p.}\oint_{(L)}\frac{f_1(\zeta)}{\zeta-z}\,d\zeta+\frac{f_1(z)}{2}$$

$$V^{\mathrm{r}}(z)=\frac{1}{2\pi i}\,\mathrm{v.p.}\oint_{(L)}\frac{f_1(\zeta)}{\zeta-z}\,d\zeta-\frac{f_1(z)}{2}$$

we can directly express the solution $u(z)$ in terms of the functions α, β and f (which is left to the reader).

Let $m>0$. Then we put

$$F(z_1)=\begin{cases} V(z_1) & \text{for } z_1 \text{ lying inside } (L)\\ z_1^m V(z_1) & \text{for } z_1 \text{ lying outside } (L)\end{cases} \tag{31}$$

This function is analytic everywhere except the points of (L) and satisfies on (L) the condition

$$F^{\mathrm{I}}(z)=F^{\mathrm{r}}(z)+f_1(z)\qquad (z\in(L))$$

Hence, $F(z)$ can be found by means of formula (30). But if an arbitrary polynomial of degree not higher than $m-1$ is added to this integral the function $V(z_1)$ defined from equalities (31) will nevertheless be equal to zero at infinity (why?). Thus we have

$$V^{\mathrm{I}}(z_1)=\frac{1}{2\pi i}\oint_{(L)}\frac{f_1(\zeta)}{\zeta-z_1}\,d\zeta+C_0+C_1z_1+\ldots+C_{m-1}z_1^{m-1}$$

$$V^{\mathrm{r}}(z_1)=\frac{1}{2\pi i z_1^m}\oint_{(L)}\frac{f_1(\zeta)}{\zeta-z_1}\,d\zeta+\frac{C_0}{z_1^m}+\frac{C_1}{z_1^{m-1}}+\ldots+\frac{C_{m-1}}{z_1}$$

where C_0, C_1, ..., C_{m-1} are arbitrary constants. By analogy with the foregoing paragraph it can be shown that there are no other solutions. Consequently, original equation (23) has an m-parameter family of solutions.

Finally, let $m<0$. Introducing again function (31) we obtain, for large $|z_1|$, Laurant's expansion

$$V^{\mathrm{r}}(z_1)=z_1^{-m}F^{\mathrm{r}}(z_1)=z_1^{|m|}\frac{-1}{2\pi i}\frac{1}{z_1}\oint_{(L)}\frac{f_1(\zeta)}{1-\frac{\zeta}{z_1}}\,d\zeta=$$

$$=-\frac{1}{2\pi i}\sum_{k=0}^{\infty}z^{|m|-1-k}\oint_{(L)}f_1(\zeta)\,\zeta^k\,d\zeta$$

But since $V^{\mathrm{r}}(\infty)$ must be equal to zero we arrive at the equalities

$$\oint_{(L)}f_1(\zeta)\,\zeta^k\,d\zeta=0\qquad (k=0,\ 1,\ \ldots,\ |m|-1) \tag{32}$$

which are necessary and sufficient for the solvability of the problem. If these equalities hold, equation (23) has exactly one solution, but if at least one of them is violated, the solution does not exist.

If we introduce the adjoint (associated) equation to equation (23) (see Sec. 3), it is readily proved that its index is equal to $-m$. Hence, for $m<0$ the homogeneous adjoint equation has $|m|$ linearly independent solutions, and it turns out that conditions (32) are nothing but the conditions of the orthogonality of the function $f(z)$ to these solutions.

For a more general equation

$$\alpha(z)\,u(z)=\frac{\beta(z)}{\pi i}\,\text{v.p.}\oint\limits_{(L)}\frac{u(\zeta)}{\zeta-z}\,d\zeta+\overset{*}{\oint\limits_{(L)}}K(z,\ \zeta)\,u(\zeta)\,d\zeta+$$
$$+f(z)\ \ (z\in(L),\ \alpha^2(z)-\beta^2(z)\neq 0) \tag{33}$$

where K is a Fredholm kernel or one with polar singularity, it is possible to combine the integral involving this kernel with f, after which the above formulas can be applied. Here, as in the passage from (19) to (20), we obtain for $u(z)$ Fredholm's equation of the second kind. However, if equation (33) has an index $m>0$ (it is calculated according to the same rule which we used for equation (23)) the nonhomogeneity term of the resultant Fredholm equation will contain m arbitrary constants. If $m<0$ the combination $\oint K u\,d\zeta+f$ should be subject to $|m|$ orthogonality conditions which are nothing but the orthogonality conditions imposed on f because u is expressed linearly in terms of f.

If the homogeneous adjoint equation of (33) is regularized in this way we can show that the Fredholm equations thus obtained are mutually adjoint, and, therefore, by Sec. 2.2, the corresponding homogeneous Fredholm equations have the same number (say, $k\geqslant 0$) of linearly independent solutions (if $k=0$ they have only a trivial solution). A more thorough analysis which we do not present here shows that for $m\geqslant 0$ the homogeneous equation (33) has $m+l$ linearly independent solutions where $0\leqslant l\leqslant k$ ($l=0$ is a typical case here) whereas the homogeneous adjoint equation to (33) has l solutions; if $m<0$ the numbers of the solutions are, respectively, equal to l and $|m|+l$. Thus, in all cases *the difference between the number of linearly independent solutions of homogeneous equation* (33) *and the corresponding number for the adjoint equation is equal to the index of equation* (33) *which does not depend on the completely continuous operator* $\int K d\zeta$ *entering into the equation*. Besides, it can be proved that *for the existence of at least one solution of nonhomogeneous equation* (33) *it is necessary and sufficient* (like in the case of Fredholm's equations) *that the*

nonhomogeneity term $f(x)$ be orthogonal to all the solutions of the homogeneous adjoint equation. (The necessity of this condition can be easily proved by means of the general scheme: if $\mathbf{Au}=\mathbf{f}$ and $\mathbf{A^*v}=\mathbf{0}$ then $(\mathbf{f}, \mathbf{v})=(\mathbf{Au}, \mathbf{v})=(\mathbf{u}, \mathbf{A^*v})=(\mathbf{u}, \mathbf{0})=0$; the proof of sufficiency is more sophisticated.)

The last two assertions also hold for nonclosed contours, for systems of singular equations of form (33) and for similar equations with any number of independent variables. (For a system of form (33) where α and β are square matrices the index is $\frac{1}{2\pi}$ times the increment of the argument of the expression $\frac{\det[\alpha(z)+\beta(z)]}{\det[\alpha(z)-\beta(z)]}$ gained as z traverses the contour (L).) Thus, the index is an essential characteristic of a nondegenerate singular integral equation; for Fredholm's equation it is always equal to zero whereas for a singular equation it may be different from zero. If we deal with a Fredholm equation, its finite-dimensional analogue is a system of linear algebraic equations in which the number of unknowns coincides with the number of equations. For a singular integral equation such an analogue is a linear system with different numbers of equations and unknowns, the difference between them playing the role analogous to that of the index of the integral equation (cf. Sec. IV.1.4; also see [107], Sec. XI.5).

The nondegeneracy of equation (33) is understood as the condition $\alpha^2(z)-\beta^2(z)\neq 0$. If this difference has zeros the situation becomes more complicated. For instance, let us suppose that $\alpha(z)+\beta(z)\equiv 0$, $\alpha(z)-\beta(z)\neq 0$. Then condition (25) readily shows that the function $f(z)$ must satisfy an infinite number of orthogonality conditions (let the reader investigate this property for the case when (L) is a circle). If these conditions are fulfilled the general solution of equation (23) will contain an infinitely many arbitrary constants because of the arbitrariness of U^r.

The addition of a singular term of a more general form $\text{v.p.}\oint\limits_{(L)}\frac{R(z,\zeta)}{\zeta-z}u(\zeta)\,d\zeta$ to the right-hand side of (33) where the kernel $R(z, \zeta)$ is continuous for $\zeta=z$ does not lead to an essential generalization. Indeed, if we write

$$\text{v.p.}\oint\limits_{(L)}\frac{R(z,\zeta)}{\zeta-z}u(\zeta)\,d\zeta=R(z,z)\,\text{v.p.}\int\limits_{(L)}\frac{u(\zeta)}{\zeta-z}d\zeta+\int\limits_{(L)}\frac{R(z,\zeta)-R(z,z)}{\zeta-z}u(\zeta)\,d\zeta$$

and combine the first of the integrals we have obtained with the

first term on the right-hand side of (33) and the second integral with the second term we again arrive at an equation of form (33).

The singular equation

$$\alpha(x)\,u(x)=\frac{\beta(x)}{2\pi}\int u(\xi)\cot\frac{\xi-x}{2}\,d\xi+\int K(x,\ \xi)\,u(\xi)\,d\xi+f(x)$$

with Hilbert's kernel $\cot\frac{\xi-x}{2}$ can be reduced to equation (33) by the substitution $z=e^{ix}$, $\zeta=e^{i\xi}$, the circle $|z|=1$ serving as the contour (L) (check it up!).

6. The Case of a Nonclosed Contour. Now we shall consider equation (23) for a nonclosed contour (L) with an initial point a and a terminal point b. Let us denote by $g(z)$ a concrete branch of the function $\operatorname{Ln}\frac{\alpha(z)+\beta(z)}{\alpha(z)-\beta(z)}$ continuous for $z\in(L)$, pass to the corresponding boundary-value problem and introduce the function

$$V(z_1)=U(z_1)\,(z_1-a)^{\mu}\,(z_1-b)^{\nu}\exp\left[\frac{-1}{2\pi i}\int\limits_{(L)}\frac{g(\zeta)}{\zeta-z_1}\,d\zeta\right] \tag{34}$$

where μ, ν are some integers which will be specified later. This is a one-valued analytic function defined throughout the z_1-plane except the points of the line (L) at which it satisfies the condition

$$V^{+}(z)=V^{-}(z)+f_1(z)\qquad(z\in(L))$$

where

$$f_1(z)=$$

$$=\frac{f(z)}{\alpha(z)-\beta(z)}\,(z-a)^{\mu}\,(z-b)^{\nu}\exp\left[-\frac{1}{2\pi i}\,\text{v.p.}\int\limits_{(L)}\frac{g(\zeta)}{\zeta-z}\,d\zeta-\frac{1}{2}\,g(z)\right] \tag{35}$$

(check it up and compare with (29)!).

Generally speaking, a function of form (35) may be nonintegrable since it may rapidly increase for $z\to a$ and $z\to b$. Consequently, to use the formula

$$V(z_1)=\frac{1}{2\pi i}\int\limits_{(L)}\frac{f_1(\zeta)}{\zeta-z_1}\,d\zeta \tag{36}$$

which is necessary for our further aims we must choose the exponents μ and ν in formula (34) in such a way that the function $f_1(z)$ is integrable. The expression in the square brackets in (35) is of the order of $\frac{1}{2\pi i}\,g(a)\ln(z-a)$ for $z\to a$, and hence the exponential expression is of the order of $|z-a|^{\operatorname{Re}\left(\frac{1}{2\pi i}\,g(a)\right)}$ (why?).

Thus, the integer μ must satisfy the inequality

$$\mu + \operatorname{Re}\left(\frac{1}{2\pi i} g(a)\right) > -1$$

On the other hand, formula (34) and the integrability condition for $U(z)$ in the vicinity of the point $z = a$ readily indicate that the above sum must be less than 1 (similarly for $U(z)$ to be finite this sum must not exceed zero). Analogously, as ν we must take an integer satisfying the inequality

$$-1 < \nu - \operatorname{Re}\left(\frac{1}{2\pi i} g(b)\right) < 1$$

Now we apply formula (36). (If $m = \mu + \nu > 0$, an arbitrary polynomial of degree not higher than $m-1$ can be added to the right side because, as is seen from (34), the resultant function $U(z_1)$ will be a solution of the boundary-value problem we deal with.) Thus, for $m > 0$ equation (23) has an m-parameter family of solutions. For $m = 0$ there is exactly one solution. Finally, for $m < 0$ the function $V(z_1)$ must be of the order of $o(z_1^m)$ for $z_1 \to \infty$; as in Sec. 5, we conclude that for this purpose the function $f(z)$ must satisfy $|m|$ orthogonality conditions obtained from the series expansion of integral (36). If these conditions hold, there is exactly one solution.

Similarly, the results of Sec. 5 can be extended to equations of form (33) with a nonclosed contour (L) and to the case when (L) consists of several nonclosed arcs.

As an example, let us derive inversion formula (15). To this end, we take the equation

$$0 = \frac{-i}{\pi i} \text{v.p.} \int_{(L)} \frac{u(\zeta)}{\zeta - z} d\zeta + F(z) \qquad ((L): \operatorname{Im} z = 0, \ -1 \leqslant z \leqslant 1)$$

Here we have

$$\alpha \equiv 0, \quad \beta \equiv -i, \quad g \equiv \pi i$$

As μ we can take the number -1 (which guarantees that the solution is finite for $z = -1$) or 0 (which increases the index but may result in an unbounded solution). Similarly, ν can be taken equal to 0 or 1. Let us put $\mu = 0$ and $\nu = 1$. Then we obtain (check it up!)

$$f_1(z) = \frac{F(z)}{i}(z-1)\exp\left[-\frac{1}{2\pi i} \text{v.p.} \int_{(L)} \frac{\pi i}{\zeta - z} d\zeta - \frac{1}{2}\pi i\right] = F(z)\sqrt{1-z^2}$$

(where the radical is understood as the arithmetic root)

$$V(z_1) = \frac{1}{2\pi i} \int_{(L)} \frac{F(\zeta)\sqrt{1-\zeta^2}}{\zeta - z_1} d\zeta + \text{const}$$

$$U(z_1) = \frac{1}{2\pi i (z_1 - 1)} \exp\left\{\frac{1}{2}\left[\ln(\zeta - z_1)\right]_{\zeta=-1}\right\} \times$$

$$\times\left[\int_{(L)}\frac{F(\zeta)\sqrt{1-\zeta^2}}{\zeta-z_1}d\zeta+\text{const}\right]$$

$$u(x)=U(x+i0)-U(x-i0)=$$

$$=\frac{1}{2\pi i(x-1)}\left\{i\sqrt{\frac{1-x}{1+x}}\left[\text{v.p.}\int_{-1}^{1}\frac{F(\xi)\sqrt{1-\xi^2}}{\xi-x}d\xi+\right.\right.$$

$$\left.+\pi iF(x)\sqrt{1-x^2}+C\right]+i\sqrt{\frac{1-x}{1+x}}\left[\text{v.p.}\int_{-1}^{1}\frac{F(\xi)\sqrt{1-\xi^2}}{\xi-x}d\xi-\right.$$

$$\left.\left.-\pi iF(x)\sqrt{1-x^2}+C\right]\right\}=-\frac{1}{\pi}\text{v.p.}\int_{-1}^{1}\sqrt{\frac{1-\xi^2}{1-x^2}}\frac{F(\xi)}{\xi-x}d\xi-\frac{C}{\pi\sqrt{1-x^2}}$$

We have thus derived formula (15). We suggest that the reader check that if we take $\mu=\nu=0$, i.e. seek a solution which is finite in the vicinity of the point $x=1$, we obtain

$$u(x)=-\frac{1}{\pi}\text{v.p.}\int_{-1}^{1}\sqrt{\frac{1+\xi}{1+x}\cdot\frac{1-x}{1-\xi}}\frac{F(\xi)}{\xi-x}d\xi$$

Also consider the case $\mu=-1$, $\nu=0$ in which there appears an additional condition of the form

$$\int_{-1}^{1}F(\xi)(1-\xi^2)^{-\frac{1}{2}}d\xi=0$$

7. Reduction to an Infinite System of Algebraic Equations. Like Fredholm's equations (Sec. 2.3), some singular integral equations can be solved numerically by reducing them directly (without regularization) to an infinite system of linear algebraic equations. For instance consider the equation

$$au(x)=b\,\text{v.p.}\int_{-1}^{1}\frac{K(x,\xi)}{\xi-x}u(\xi)\,d\xi+f(x)\qquad(-1\leqslant x\leqslant 1)\quad(37)$$

Performing the change of variables

$x=\cos\theta$, $\xi=\cos\varphi$, $u(x)=v(\theta)$, $f(x)=F(\theta)$, $K(x,\xi)=L(\theta,\varphi)$

we rewrite (37) in the form

$$av(\theta)=b\,\text{v.p.}\int_{0}^{\pi}\frac{L(\theta,\varphi)}{\cos\varphi-\cos\theta}v(\varphi)\sin\varphi\,d\varphi+F(\theta)\qquad(0\leqslant\theta\leqslant\pi)\quad(38)$$

Now the function $v(\theta)$ should be expanded into a series with respect to a complete system of functions. After this expansion has been

substituted into (38), we obtain a system of linear algebraic equations in the coefficients of the expansion. For example, let us take the following expansions:

$$\left.\begin{aligned} v(\theta) &= \sum_{j=0}^{\infty}{}' v_j \cos j\theta, \quad F(\theta) = \sum_{j=0}^{\infty} F_j \cos j\theta \\ L(\theta, \varphi) &= \sum_{j=0}^{\infty}\sum_{k=1}^{\infty} L_{jk} \cos j\theta \sin k\varphi \end{aligned}\right\} \tag{39}$$

Here we have

$$L(\theta, \varphi)\, v(\varphi) = \sum_{j=0}^{\infty}\sum_{k=1}^{\infty}\sum_{s=0}^{\infty} L_{jk} v_s \cos j\theta \sin k\varphi \cos s\varphi =$$

$$= \frac{1}{2}\sum_{j=0}^{\infty}\sum_{k=1}^{\infty}\sum_{s=0}^{\infty} L_{jk} v_s \cos j\theta \left[\sin (k+s)\varphi + \sin (k-s)\varphi\right]$$

and therefore, by virtue of the formulas given at the end of Sec. II.3.5, we receive

$$\text{v.p.} \int_0^{\pi} \frac{L(\theta, \varphi)}{\cos\varphi - \cos\theta} v(\varphi) \sin\varphi\, d\varphi =$$

$$= -\frac{\pi}{2}\sum_{j=0}^{\infty}\sum_{k=1}^{\infty} L_{jk} \cos j\theta \left(\sum_{s=0}^{\infty} v_s \cos (k+s)\theta + \sum_{s=0}^{\infty}{}' v_s \cos (k-s)\theta\right) \tag{40}$$

where the prime means that the term with $k = s$ is omitted. For the sake of convenience, let us extend the summation in (40) to all values of the indices ranging from $-\infty$ to ∞ by putting the coefficients which do not enter into (39) equal to zero. Then, the right-hand side of (40) is transformed to

$$-\frac{\pi}{4}\sum_{j}\sum_{k}\left\{\sum_{s} L_{jk} v_s \left[\cos (k+s+j)\theta + \cos (k+s-j)\theta\right] + \right.$$

$$\left. + \sum_{s}{}' L_{jk} v_s \left[\cos (k-s+j)\theta + \cos (k-s-j)\theta\right]\right\}$$

Now we break up this expression into four sums corresponding to the cosines with different arguments and pass to the summation with respect to k, s and $r = k+s+j$ in the first sum, with respect to k, s and $r = k+s-j$ in the second sum, etc. Combining the results we obtain

$$-\frac{\pi}{4}\sum_{r}\sum_{k}\left[\sum_{s} (L_{r-k-s,\,k} + L_{k+s-r,\,k})\, v_s + \right.$$

$$\left. + \sum_{s}{}' (L_{r-k+s,\,k} + L_{k-s-r,\,k})\, v_s\right] \cos r\theta$$

Replacing the index r by j, substituting the resultant expression into (38) and equating the coefficients in the cosines we obtain the expressions

$$av_0 = -\frac{\pi}{4} b \sum_{s=0}^{\infty} \left[\sum_{k=1}^{\infty} L_{k+s,\,k} + \sum_{k=1}^{\infty}{}' (L_{-k+s,\,k} + L_{k-s,\,k}) \right] v_s + F_0$$

$$av_j = -\frac{\pi}{4} b \sum_{s=0}^{\infty} \left[\sum_{k=1}^{\infty} (L_{j-k-s,\,k} + L_{k+s-j,\,k} + L_{k+s+j,\,k}) + \right.$$

$$\left. + \sum_{k=1}^{\infty}{}' (L_{j-k+s,\,k} + L_{k-s-j,\,k} + L_{-j-k+s,\,k} + L_{k-s+j,\,k}) \right] v_s + F_j$$

$$(j = 1, 2, 3, \ldots)$$

where the coefficients $L_{p,\,k}$ for $p < 0$ are equal to zero. This is an infinite linear system with an infinite coefficient matrix which can sometimes be solved with the help of techniques discussed in Sec. 2.3 but it should be taken into account that in contrast to Sec. 2.3 the elements of the principal diagonal of the coefficient matrix (and also of parallel diagonals) have nonzero limits. If the principal diagonal coefficients are dominant, the corresponding terms can be transposed to the left-hand side, after which the iteration scheme can be applied; in addition, we can also transpose the terms corresponding to the elements of the nearest parallel diagonals and then apply the scheme of continued fractions with iterations, etc.

§ 6. NONLINEAR INTEGRAL EQUATIONS

The theory of *nonlinear integral equations* is by far not so complete as the theory of linear equations. There are many theorems on the solvability and on the properties of solutions proved for various classes of nonlinear equations (in particular, see [67], [69], [71], [133], and [134]) and a number of approaches to a numerical solution of nonlinear integral equations (though their practical application is often connected with substantial difficulties). Here we shall only discuss some of these approaches.

1. Reduction to Finite Equations. There is a special class of integral equations similar to linear equations with degenerate kernels (Sec. 2.1) which are directly reducible to finite equations. For instance, let us consider the equation

$$u(x) = \int_a^b \sum_{j=1}^{n} \Phi_j(x)\, \Psi_j(\xi,\ u(\xi))\, d\xi + f(x) \qquad (a \leqslant x \leqslant b) \qquad (1)$$

Putting

$$\int_a^b \Psi_j(\xi, u(\xi))\, d\xi = C_j \qquad (j = 1, 2, \ldots, n) \tag{2}$$

we obtain

$$u(x) = \sum_{j=1}^{n} C_j \Phi_j(x) + f(x)$$

Substituting this expression into (2) we arrive at the system of finite nonlinear equations

$$\int_a^b \Psi_j\left(\xi, \sum_{j=1}^{n} C_j \Phi_j(\xi) + f(\xi)\right) d\xi = C_j \qquad (j = 1, 2, \ldots, n) \tag{3}$$

which is completely equivalent to equation (1). Various methods of numerical solution of systems of finite equations can be applied to system (3), and in some simpler cases even the exact solution can be found.

The simple example below demonstrates some specific features of nonlinear equations. Let us take the equation

$$u(x) = -\frac{\lambda}{2} \int_0^1 x\left[u^3(\xi) + 3u^2(\xi) + 4u(\xi)\right] d\xi \tag{4}$$

involving the parameter λ and limit ourselves to investigating its real solutions. According to the above we put

$$\int_0^1 \left[u^3(\xi) + 3u^2(\xi) + 4u(\xi)\right] d\xi = C \text{ whence } u(x) = -\frac{\lambda C}{2} x$$

After the substitution we obtain the following relationship between C and λ:

$$\lambda^3 C^3 - 8\lambda^2 C^2 + 32\lambda C + 32C = 0 \tag{5}$$

Equation (5) specifies a curve (consisting of two branches) in the λC-plane which is shown in heavy line in Fig. 118. We see that for $-\infty < \lambda < -2$ and for $0 \leqslant \lambda < \infty$ equation (4) has only a zero solution, and for $-2 < \lambda < 0$ it has two additional nonzero solutions. Therefore, if, by analogy with the linear case, we consider a given λ to be a characteristic value when equation (4) has at least one nonzero solution, there appears a whole interval of characteristic values; however, to each such value there corresponds a finite number of solutions (e.g. three solutions in this particular case) while in the case of linear equations we have an infinite number of solutions. The value $\lambda = -1$ is of particular interest here. Suppose that we consider a solution which is continuous in

λ when λ increases beginning with large negative values. Then for $\lambda < -1$ this solution is identically equal to zero while for $\lambda > -1$ there are two branches: one remaining zero and the other different from zero and receding to infinity as $\lambda \to -0$. Thus, $\lambda = -1$ is a **bifurcation point** of the given integral equation. If the parameter λ varies in the opposite direction the bifurcating solution is equal to zero for large $\lambda > -1$ and branches off at the point $\lambda = -1$ into two solutions one of which remains zero while the other is nonzero and disappears (or more precisely, becomes imaginary) as $\lambda \to -2 - 0$.

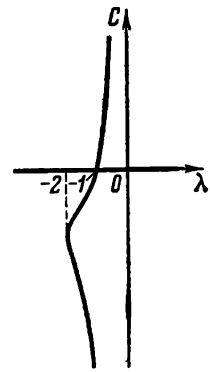


Fig. 118

It should be stressed that the fact that for certain λ the nonlinear equation (4) has only a zero solution does not mean that the corresponding nonhomogeneous equation (with an additional term $f(x)$ on the right-hand side) has exactly one solution because none of the Fredholm theorems is applicable here.

The direct reduction to finite equations can also be used for equations of the form

$$u(x) = F\left(x, \int_a^b \Phi_1(x)\,\Psi_1(\xi,\ u(\xi))\,d\xi,\ \int_a^b \Phi_2(x)\,\Psi_2(\xi,\ u(\xi))\,d\xi,\ \ldots\right)$$

which are more general than (1).

The exact reduction to finite equations is possible only in some special cases but an approximate reduction can always be used and is one of the main methods of numerical solution and investigation of nonlinear equations. For instance, consider the **Uryson equation** (P.S. Uryson (1898-1924), a Soviet mathematician)

$$u(x) = \int_a^b f(x, \xi, u(\xi))\,d\xi \qquad (a \leqslant x \leqslant b) \tag{6}$$

(we can also consider a many-dimensional analogue of this equation). We shall try to construct an approximate solution of the form $u = \varphi(x, C_1, C_2, \ldots, C_n)$ where, as usual, the function φ should be chosen so that it yields an accurate approximation to the exact solution for a small number of parameters. Substituting this expression into (6) we obtain the discrepancy

$$\varphi(x, C_1, C_2, \ldots, C_n) - \int_a^b f(x, \xi, \varphi(\xi, C_1, C_2, \ldots, C_n))\,d\xi \qquad (a \leqslant x \leqslant b)$$

which can then be minimized by means of the method of moments, by the method of least squares or by the collocation method (e.g.

see [107], Sec. XV.29 where these methods are considered in connection with differential equations).

2. Iteration Method. The iteration method described in Sec. 2.5 is also applicable to some classes of nonlinear equations. These are, first of all, Volterra's nonlinear equations of the second kind

$$\left.\begin{aligned} u(x) &= \int_a^x f(x, \xi, u(\xi))\, d\xi + \varphi(x) \\ u(x) &= F\left(x, \int_a^x K(x, \xi)\, u(\xi)\, d\xi\right) \\ u(x) &= \varphi(x) + \int_a^x K(x, \xi)\, u(\xi)\, d\xi + \\ &\quad + \int_a^x \int_a^x L(x, \xi, \eta)\, u(\xi)\, u(\eta)\, d\xi\, d\eta \end{aligned}\right\} \qquad (7)$$

and also some other types of nonlinear equations whose common feature is that they are all resolved with respect to $u(x)$ and their right-hand sides include u under the sign of integration with the variable upper limit x. The possibility of solving an equation of type (7) by means of a step-by-step procedure with small steps along the x-axis which was discussed in Sec. 2.9 and the possibility of using linearization on each step account for the fact that, as a rule, each equation of this type has a single solution which can be obtained with the aid of the standard iteration scheme beginning with any zeroth approximation.

In connection with the discussion of Volterra's equations it should be noted that if in the special types of equations considered in Sec. 1 we replace $\int_a^b$ by $\int_a^x$ it is easy to pass to an initial-value problem for ordinary differential equations which can be solved by numerical integration. For instance, if we take the equation

$$u(x) = \int_a^x \sum_{j=1}^n \Phi_j(x)\, \Psi_j(\xi, u(\xi))\, d\xi + f(x) \qquad (8)$$

and put

$$\int_a^x \Psi_j(\xi, u(\xi))\, d\xi = v_j(x)$$

the problem reduces to the system of differential equations

$$v_j' = \Psi_j\left(x, \sum_{j=1}^n \Phi_j(x)\, v_j + f(x)\right) \qquad (j = 1, 2, \ldots, n)$$

which, for initial conditions

$$v_j(a)=0 \qquad (j=1, 2, \ldots, n)$$

is completely equivalent to original equation (8).

The iteration method is also applicable to equations with small nonlinear kernels. For instance, take the equation

$$u(x)=\mu\int_a^b f(x, \xi, u(\xi))\,d\xi+\varphi(x) \tag{9}$$

where $|\mu|$ is sufficiently small. Successive approximations are naturally determined by the formula

$$u_{n+1}(x)=\mu\int_a^b f(x, \xi, u_n(\xi))\,d\xi+\varphi(x) \qquad (n=0, 1, 2, \ldots)$$

The equality

$$u_{n+1}(x)-u_n(x)=\mu\int_a^b [f(x, \xi, u_n(\xi))-f(x, \xi, u_{n-1}(\xi))]\,d\xi \tag{10}$$

shows that for small $|\mu|$ these approximations are convergent with the rate of a geometric progression. In practical computations the applicability of the iteration scheme for a given μ can be tested by comparing subsequent approximations.

Some explicit conditions for convergence of iterations can be obtained on the basis of the contraction mapping principle (Sec. 2.6); however, they are rarely used in problem-solving practice. Let us consider the iteration process in the space $C[a, b]$ and suppose that the inequality

$$|f'_u(x, \xi, u)|\leqslant K(x, \xi) \tag{11}$$

holds for all values of u or at least for all values involved in the iteration process. Denoting by $\rho(u, v)$ the uniform deviation of the function $u(x)$ from $v(x)$ (which is the distance between u and v in the space $C[a, b]$) and applying the formula of finite increments (e.g. see [107], Sec. V.4) we derive from (10) the inequality

$$|u_{n+1}(x)-u_n(x)|\leqslant|\mu|\int_a^b K(x, \xi)\,\rho(u_n, u_{n-1})\,d\xi$$

whence

$$\rho(u_{n+1}, u_n)\leqslant|\mu|\max_x\int_a^b K(x, \xi)\,d\xi\cdot\rho(u_n, u_{n-1})$$

By the contraction mapping principle it follows that the iterations converge if

$$|\mu| < \left(\max_x \int_a^b K(x, \xi)\, d\xi\right)^{-1} \tag{12}$$

(In particular, if condition (11) holds for all u equation (9) has a unique solution for every value of μ satisfying inequality (12), but if (11) holds only for the values of u used in the iteration process the uniqueness can be violated.) Inequality (12) provides only a sufficient condition for convergence which in fact may take place for a wider range of variation of μ, in particular, when $|\mu|$ is less than the least characteristic value of the kernel K. A more thorough analysis shows that for sufficiently small values of $|\mu|$ the iteration process is convergent even if

$$\max_x \int_a^b K(x, \xi)\, d\xi = \infty$$

3. Small-Parameter Method. In the case of small μ equation (9) admits of the application of the small-parameter method (e.g. see [107], Sec. V.5). This method can be used for any linear equation perturbed by a small nonlinear term. We shall demonstrate the application of the method by the **Lyapunov-Lichtenstein equation**

$$u(x) = f(x) + \lambda \int_a^b K_1(x, \xi)\, u(\xi)\, d\xi + \\ + \mu \int_a^b \int_a^b K_2(x, \xi, \mu)\, u(\xi)\, u(\eta)\, d\xi\, d\eta \tag{13}$$

(to which similar terms with K_3, K_4, etc. can also be added) under the assumption that λ is not a characteristic value of the kernel K_1 (which can be a Fredholm kernel or one with a polar singularity), μ being a small parameter. Substituting the expansion

$$u(x) = u_0(x) + \mu u_1(x) + \mu^2 u_2(x) + \ldots \tag{14}$$

into (13) and equating the coefficients in like powers of μ we arrive at the equalities

$$u_0(x) = f(x) + \lambda \int_a^b K_1(x, \xi)\, u_0(\xi)\, d\xi \tag{15}$$

$$u_1(x) = \lambda \int_a^b K_1(x, \xi)\, u_1(\xi)\, d\xi + \int_a^b \int_a^b K_2(x, \xi, \eta)\, u_0(\xi)\, u_0(\eta)\, d\xi\, d\eta \tag{16}$$

$$u_2(x) = \lambda \int_a^b K_1(x, \xi) u_2(\xi)\, d\xi +$$

$$+ \int_a^b \int_a^b K_2(x, \xi, \eta) [u_0(\xi) u_1(\eta) + u_1(\xi) u_0(\eta)]\, d\xi\, d\eta \qquad (17)$$

etc. We see that the construction of every coefficient in expansion (14) reduces to solving a linear equation with one and the same kernel K_1 and a nonhomogeneity term determined by the foregoing coefficients. By the hypothesis about λ this construction is possible; it becomes particularly simple if the resolvent kernel of the kernel K_1 is known.

Now let λ be a characteristic value of the kernel K_1 with which there are associated k linearly independent eigenfunctions. Then, by virtue of (15), the function $f(x)$ must satisfy k conditions expressing the orthogonality to the eigenfunctions of the adjoint equation; if these conditions hold, the solution $u_0(x)$ of equation (15) contains arbitrary constants $C_1, C_2, \ldots, C_k$. Substituting this solution into equation (16) and taking advantage of the orthogonality conditions of the nonhomogeneity term of the equation to the eigenfunctions of the adjoint equation, we obtain a system of k quadratic equations in $C_1, C_2, \ldots, C_k$ from which we find these constants (for an original integral equation of some other type this procedure may lead to a system of k finite equations of a more general structure). From the latter conditions we find C_j $(j = 1, 2, \ldots, k)$ and then obtain the general solution of equation (16) with arbitrary constants $C_1', C_2', \ldots, C_k'$. Substituting this solution into the nonhomogeneity term of equation (17) we find these constants from the corresponding orthogonality conditions (the equations for $C_1', C_2', \ldots, C_k'$ turn out to be linear, i.e. these constants can be easily found; the same refers to further terms of expansion (14)). Going on in this way we can construct any number of terms of expansion (14).

In equation (13) the functions f, K_1 and K_2 and even the limits of integration can also depend on μ; in this case, after all these quantities are expanded in powers of μ, the above method can be used in a similar way.

4. Application of the Theory of Symmetric Kernels. Consider the **Hammerstein equation**

$$u(x) = \int_a^b K(x, \xi)\, \Phi(\xi, u(\xi))\, d\xi + f(x) \qquad (18)$$

with a symmetric kernel K.

For our further aims we need the inequality

$$\left\| \int_a^b K(x, \xi) v(\xi) d\xi \right\| \leqslant \frac{1}{|\lambda_1|} \| v(x) \| \tag{19}$$

where the norm is taken in the space $L_2[a, b]$ and λ_1 is the least in its modulus characteristic value of the kernel K. To prove (19) we use (3.8) and thus obtain

$$\int_a^b \left[\int_a^b K(x, \xi) v(\xi) d\xi \right]^2 dx = \int_a^b \left[\sum_j \frac{(v, \varphi_j)}{\lambda_j} \varphi_j(x) \right]^2 dx =$$

$$= \sum_j \frac{(v, \varphi_j)^2}{\lambda_j^2} \leqslant \frac{1}{|\lambda_1|^2} \sum_j (v, \varphi_j)^2 \leqslant \frac{1}{|\lambda_1|^2} \| v(x) \|^2$$

(check it up!).

Now, taking equation (18) we can readily prove that if

$$\Phi'_u(\xi, u) \leqslant \alpha < |\lambda_1| \tag{20}$$

the equation has exactly one solution which can be found with the help of the iteration method. Actually, the right-hand side of (18) specifies a mapping Bu of the space $L_2[a, b]$ into itself (it is supposed that $f \in L_2$). But we have

$$\| Bu_1 - Bu_2 \| = \| \int_a^b K(x, \xi) [\Phi(\xi, u_1(\xi)) - \Phi(\xi, u_2(\xi))] d\xi \| \leqslant$$

$$\leqslant \frac{1}{|\lambda_1|} \| \Phi(x, u_1(x)) - \Phi(x, u_2(x)) \| \leqslant \frac{\alpha}{|\lambda_1|} \| u_1 - u_2 \|$$

and hence our assertion is implied by the contraction mapping principle (Sec. 2.6).

Let us apply the result to the theory of oscillations of a **simple pendulum** (which is a particle suspended by a weightless inextensible cord). The differential equation of plane forced oscillations without friction has the form

$$\ddot{\varphi} + \frac{g}{l} \sin \varphi = \frac{1}{ml} F(t) \tag{21}$$

where φ is the angular deviation from the vertical, g is the acceleration of gravity, l is the length of the cord, m is the mass of the pendulum, and $F(t)$ is the tangential external force. Let us suppose that $F(t)$ and the corresponding solution $\varphi(t)$ are odd periodic functions of time t with a given period T. Then it is sufficient to construct the solution on the interval $0 \leqslant t \leqslant \frac{T}{2}$ for the boundary conditions

$$\varphi(0) = 0, \ \varphi\left(\frac{T}{2}\right) = 0 \tag{22}$$

after which it can be periodically extended in odd fashion to obtain the solution for all t.

The solution of the equation $\ddot{\varphi} = f(t)$ for boundary conditions (22) has the form

$$\varphi(t) = \int_0^{\frac{T}{2}} G(t, \tau) f(\tau) d\tau$$

(e.g. see [107], Sec. XV.16) where the symmetric kernel G is defined by the formula

$$G(t, \tau) = \begin{cases} -\frac{(T-2\tau)t}{T} & \text{for } 0 \leqslant t \leqslant \tau \leqslant \frac{T}{2} \\ -\frac{(T-2t)\tau}{T} & \text{for } 0 \leqslant \tau \leqslant t \leqslant \frac{T}{2} \end{cases}$$

Applying this result to equation (21) we see that this equation taken with boundary conditions (22) is equivalent to the integral equation

$$\varphi(t) = -\frac{g}{l} \int_0^{\frac{T}{2}} G(t, \tau) \sin \varphi(\tau) d\tau + \frac{1}{ml} \int_0^{\frac{T}{2}} G(t, \tau) F(\tau) d\tau \qquad (23)$$

The characteristic values of the kernel G are equal to the eigenvalues of the equation $\ddot{\varphi} + \lambda\varphi = 0$ (with boundary conditions (22)) taken with the opposite sign (why?). Therefore, the least of the moduli of the characteristic values of G is equal to $\lambda_1 = -\left(\frac{2\pi}{T}\right)^2$. Now, applying the general condition (20) we see that if the inequality

$$\frac{gT^2}{l} < 4\pi^2 \qquad (24)$$

is fulfilled the problem under consideration possesses exactly one solution which can be found by iteration.

5. Application of Fixed-Point Theorems. As was mentioned in Secs. 2.6. and 4, the determination of solutions of many equations can be reduced to finding a fixed point of an appropriately chosen mapping of a point set into itself. The contraction mapping principle is one of the main methods for establishing the existence of such a fixed point. Another method of finding a fixed point in a finite-dimensional case is based on the application of the Bohl-Brouwer theorem (Sec. IV.4.8). The following consequence of this theorem is sometimes resorted to: *if there is a continuous mapping of a bail* (P) *lying in a finite-dimensional Euclidean space* (E) *into* (E) *under which none of the boundary points of* (P) *moves away*

from the centre of the ball in a radial direction the mapping possesses at least one fixed point. (How can this proposition be derived from the Bohl-Brouwer theorem?) There are also some stronger theorems of this kind.

Such theorems are of topological character because they hold for all topologically equivalent sets. (See the end of Sec. II.1.7.) For instance, the Bohl-Brouwer theorem holds for a horse-shoe-shaped figure but it is inapplicable to an annulus (why?). That is why the methods with the help of which we can determine the existence of a solution of a problem by applying fixed-point theorems or other similar theorems are called topological. (On these methods we refer the reader to [69].) Topological methods often prove applicable in some situations when other methods fail. Unfortunately, as a rule, topological methods are not constructive. They guarantee the existence of at least one solution (which is of course very important) but give no indication as to the number of such solutions and the way they can be constructed; for this purpose other methods must be used. (It should be noted that the contraction mapping principle bears a metric character and is not topological since the notion of a contraction mapping is not topological.)

For an infinite-dimensional Banach space instead of the Bohl-Brouwer theorem we usually use the **Schauder theorem** which is formulated in exactly the same way but involves an additional requirement that the mapping $\mathbf{x} \to \mathbf{Ax}$ is not only continuous but completely continuous, i.e. the image of any bounded set under this mapping must be compact (see Sec. 2.11). (The proof of the Schauder theorem reduces to approximating the given mapping with a sequence of finite-dimensional mappings, applying the Bohl-Brouwer theorem to these mappings and using the compactness of the approximating sequence of fixed points thus obtained.) The above consequence of the Bohl-Brouwer theorem can also be applied to this mapping $\mathbf{x} \to \mathbf{Ax}$ since its conditions hold in these circumstances. There are some other similar and stronger propositions.

Let us demonstrate the application of the theory of fixed points to equation (18). Suppose that the function $\Phi(\xi, u)$ is continuous and bounded for all ξ, u $(a \leqslant \xi \leqslant b, \ -\infty < u < \infty)$, say $|\Phi(\xi, u)| \leqslant \Phi_0$. Denote the right-hand side of (18) as $\mathbf{Au}$; then from Sec. 2.11 we readily conclude that A is a completely continuous operator. Besides, from inequality (19) it follows that for all $\mathbf{u} \in L_2[a, b]$ we have

$$\|\mathbf{Au}\| \leqslant \frac{1}{|\lambda_1|} \Phi_0 \sqrt{b-a} + \|\mathbf{f}\| \tag{25}$$

Hence, if we take, as set P, a ball in $L_2[a, b]$ with centre at the origin and radius equal to the right member of (25) all the con-

ditions of the Schauder theorem are fulfilled. Thus, for any bounded continuous function Φ equation (18) has at least one solution. In particular, this applies to equation (23), i.e. the problem discussed at the end of Sec. 4 has at least one (not necessarily unique) solution even if inequality (24) is not fulfilled.

A similar result is obtained if the function $\Phi(\xi, u)$ satisfies the weaker condition

$$|\Phi(\xi, u)| \leqslant k|u| + \Phi_0(\xi) \text{ where } k < \frac{1}{|\lambda_1|}, \ \Phi_0(\xi) \in L_2[a, b] \quad (26)$$

(prove it!).

Let us suppose, in addition, that the kernel K is positive definite (Sec. 3.5) and that $f(x) \equiv 0$. In this case we can prove a still stronger proposition if we consider the mapping $\mathbf{u} \to \mathbf{Au}$ in the ball $\|\mathbf{u}\| \leqslant R$ where R is sufficiently large and apply the infinite-dimensional analogue of the consequence of the Bohl-Brouwer theorem given at the beginning of this section. We shall suppose that the function $\Phi(\xi, u)$ is continuous and satisfies the inequality

$$|\Phi(\xi, u)| \leqslant C|u| + \Phi_0(\xi) \text{ where } \Phi_0(\xi) \in L_2[a, b] \quad (27)$$

C being a constant; it can be shown that under this condition the mapping A is completely continuous. If a point of the sphere $\|\mathbf{u}\| = R$ moves away from the centre in a radial direction this means that there is a function $u(x)$ for which

$$\|\mathbf{u}\| = R, \ \int_a^b K(x, \xi) \Phi(\xi, u(\xi)) d\xi = (1+\alpha) u(x) \qquad (\alpha > 0) \quad (28)$$

Let us put

$$\int_a^b u(x) \varphi_j(x) dx = u_j, \ \int_a^b \Phi(x, u(x)) \varphi_j(x) dx = v_j \qquad (j = 1, 2, 3, \ldots)$$

where $\varphi_1, \varphi_2, \ldots$ is the system of eigenfunctions of the kernel K. Then it readily follows from (27) that

$$v_j = (1+\alpha) \lambda_j u_j \text{ whence } u_j v_j = (1+\alpha) \lambda_j u_j^2 \geqslant (1+\alpha) \lambda_1 u_j^2$$

Summing both sides of the latter inequality with respect to j we obtain (check it up!)

$$(1+\alpha) \lambda_1 \int_a^b u^2(x) dx \leqslant \int_a^b u(x) \Phi(x, u(x)) dx$$

For the above consequence of the Bohl-Brouwer theorem to be applicable we must impose a condition guaranteeing that the last relation does not hold for large $\|\mathbf{u}\|$. To this end it is sufficient to require that inequality (26) should be fulfilled when $\Phi(\xi, u)$

and u have the same sign (why?). In other words, it is sufficient to impose the condition

$$\left.\begin{aligned} \Phi(\xi, u) &\leqslant ku + \Phi_0(\xi) \text{ for } u \geqslant 0 \\ \Phi(\xi, u) &\geqslant ku - \Phi_0(\xi) \text{ for } u \leqslant 0 \end{aligned}\right\} \tag{29}$$

where $k < \frac{1}{|\lambda_1|}$ and $\Phi_0(\xi) \in L_2[a, b]$. Thus, if inequalities (27) and (29) are fulfilled, equation (18) with positive definite kernel K and $f \equiv 0$ has at least one solution.

It can also be easily proved that the result we have obtained is valid for a positive semidefinite kernel K; to this end the kernel K should be approximated with positive definite kernels.

Here we have confined ourselves to simple results concerning the application of fixed-point theorems. For more detail we refer the reader to special monographs.

6. Variational Methods. For numerical solution of linear and nonlinear integral equations we can apply variational methods. (These methods are also used to investigate general questions on solvability of nonlinear equations (see [69], [133]).)

It is sometimes possible to write the given equation in the form of Euler's equation for an appropriately chosen functional, which makes it possible to apply the Ritz method. For example, it can easily be shown that equation (2.16) with symmetric kernel is Euler's equation for the functional

$$I\{u\} = \int_a^b u^2(x)\,dx - \lambda \int_a^b \int_a^b K(x, \xi)\, u(x)\, u(\xi)\,dx\,d\xi - 2\int_a^b f(x)\, u(x)\,dx$$

Besides, it is always advisable to try to use the method of least squares (Sec. VI.4.7).

The reduction to Euler's equation is sometimes more complicated. For example, let us consider the Hammerstein equation

$$\alpha u(x) = \int_a^b K(x, \xi) f(\xi, u(\xi))\,d\xi \qquad (\alpha = \text{const}) \tag{30}$$

with a positive semidefinite symmetric kernel K whose eigenvalues λ_j and eigenfunctions $\varphi_j(x)$ are known. Let us introduce the kernel $L(x, \xi) = \sum_j \varphi_j(x)\, \varphi_j(\xi)\, \lambda_j^{-\frac{1}{2}}$ whose first iteration is K and the function $F(\xi, u) = \int_{u_0}^{u} f(\xi, s)\,ds$. Then Euler's equation for the functional

$$I\{v\} = \frac{\alpha}{2}\int_a^b v^2(x)\,dx - \int_a^b F\left(x, \int_a^b L(x, \xi)\, v(\xi)\,d\xi\right)dx \tag{31}$$

has the form (check it up!)

$$\alpha v(x) = \int_a^b L(x, \xi) f\left(\xi, \int_a^b L(\xi, \eta) v(\eta) d\eta\right) d\xi$$

Now, putting

$$u(x) = \int_a^b L(x, \xi) v(\xi) d\xi \tag{32}$$

we readily derive (30). Consequently, after an approximation for a stationary value of functional (31) has been determined we obtain, by means of (32), an approximate solution of equation (30).

We can also use an infinite sequence of coordinate functions (see Sec. VI.4.3). Let us again consider equation (30) without requiring that the kernel K should be positive-semidefinite. As the function $u(x)$ is sourcewise representable with the aid of the kernel (Sec. 3.3) for $\alpha \neq 0$ we can write the expansion

$$u(x) = \sum_j u_j \varphi_j(x)$$

(for $\alpha = 0$ we must additionally require that the system of eigenfunctions of the kernel K should be complete). After the substitution of this expression into (30) we multiply both sides scalarly by φ_j and thus obtain the system of finite equations

$$\alpha u_j = \frac{1}{\lambda_j} \int_a^b f\left(\xi, \sum_k u_k \varphi_k(\xi)\right) \varphi_j(\xi) d\xi \qquad (j = 1, 2, 3, \ldots)$$

which is equivalent to (30) and is nothing but the system of stationarity conditions for the function

$$\Phi(u_1, u_2, u_3, \ldots) = \alpha \sum_j \lambda_j u_j^2 - 2 \int_a^b F\left(\xi, \sum_j u_j \varphi_j(\xi)\right) d\xi$$

(check it up!). A stationary point can be approximately found by truncation, i.e. by means of an approximate passage to a finite number of variables. If the stationary point is a point of minimum we can use the gradient method or the method of steepest descent.

7. Equations with Parameters. It often occurs that a given integral equation contains one or several parameters and that it is necessary to investigate the variation of a solution of the equation when the parameters are changed. We shall demonstrate some simpler results of this kind by the example of the equation

$$u(x) = \int_a^b f(x, \xi, u(\xi), \gamma) d\xi \tag{33}$$

with one parameter γ though these results are directly extended to equations of more general types with any number of parameters. As before, all the quantities involved are supposed to be real.

Assume that for a value $\gamma=\gamma_0$ of the parameter a solution $u_{\gamma_0}(x)$ of equation (33) is known (the equation can also have other solutions for $\gamma=\gamma_0$). If γ is given a small variation, then, generally speaking, the variation of the solution $u=u_\gamma(x)$ is also small (except some singular cases which will be discussed later). Differentiating both sides of equation (33) with respect to γ we obtain

$$\left.\frac{\partial u_\gamma(x)}{\partial\gamma}\right|_{\gamma=\gamma_0}=\int_a^b f'_u(x,\ \xi,\ u_{\gamma_0}(\xi),\ \gamma_0)\left.\frac{\partial u_\gamma(\xi)}{\partial\gamma}\right|_{\gamma=\gamma_0}d\xi+$$

$$+\int_a^b f'_\gamma(x,\ \xi,\ u_{\gamma_0}(\xi),\ \gamma_0)\,d\xi$$

We see that the derivative $\left.\frac{\partial u_\gamma}{\partial\gamma}\right|_{\gamma=\gamma_0}$ is determined by Fredholm's linear equation of the second kind with the kernel

$$f'_u(x,\ \xi,\ u_{\gamma_0}(\xi),\ \gamma_0) \tag{34}$$

By virtue of Sec. 2.2, this derivative is determined uniquely if for kernel (34) (taken without an additional parameter λ) the first Fredholm alternative takes place, i.e. if $\lambda=1$ is not a characteristic value of this kernel. As in the theorem on the existence of an implicit function (e.g. see [107], Sec.IX.13 and XII.3), it can be shown that for small values of $|\gamma-\gamma_0|$ this condition is also sufficient for equation (33) to have a uniquely determined solution $u_\gamma(x)$ continuously depending on γ and equal to $u_{\gamma_0}(x)$ for $\gamma=\gamma_0$.

After the expansion

$$u_\gamma(x)=\varphi_0(x)+(\gamma-\gamma_0)\,\varphi_1(x)+(\gamma-\gamma_0)^2\,\varphi_2(x)+\ldots$$

has been substituted into both sides of (33) it can easily be verified that $\varphi_0(x)=u_{\gamma_0}(x)$, $\varphi_1(x)=\left.\frac{\partial u_\gamma(x)}{\partial\gamma}\right|_{\gamma=\gamma_0}$ and that each of the subsequent coefficients $\varphi_k(x)$ is determined by Fredholm's linear integral equation of the second kind with the same kernel (34) and a nonhomogeneity term expressed in terms of the coefficients $\varphi_0(x)$, $\varphi_1(x)$, ..., $\varphi_{k-1}(x)$ that have already been found. Therefore, in principle, we can find in succession any number of coefficients although if the resolvent kernel of kernel (34) is unknown this is practically by far not simple.

To find an approximate solution of equation (33) for small $|\gamma-\gamma_0|$ we can also rewrite it in the form

$$u(x)-\int_a^b f'_u(x,\ \xi,\ u_{\gamma_0}(\xi),\ \gamma_0)\,u(\xi)\,d\xi=$$

$$=\int_a^b [f(x,\ \xi,\ u(\xi),\ \gamma)-f'_u(x,\ \xi,\ u_{\gamma_0}(\xi),\ \gamma_0)\,u(\xi)]\,d\xi$$

and then apply the method of successive approximations with $u_{\gamma_0}(x)$ as initial approximation. Since for $\gamma=\gamma_0$, $u=u_{\gamma_0}(x)$ the variation of the right-hand side is equal to zero (check it up!) it remains small for the values of γ and u which are, respectively, close to γ_0 and $u_{\gamma_0}(x)$; this guarantees the convergence of the iterations (e.g. cf. [107], Sec. V.3).

The solution can be continuously extended with respect to the parameter γ provided that it does not recede to infinity for a finite value of γ and that in this process it does not take on values for which the function f is discontinuous or for which kernel (34) has $\lambda=1$ as its characteristic value.

8. Bifurcation of Solutions. Suppose that $\lambda=1$ is a characteristic value of kernel (34). We shall consider only the basic case when to this value there corresponds only one eigenfunction $u_1(x)$ (defined to within a scalar factor) and, therefore, only one eigenfunction $v_1(x)$ of the adjoint equation; we shall consider these functions to be known (theoretically, they can be determined exactly but practically only approximately). Let us consider the auxiliary equation

$$u(x)=\int_a^b [f(x,\ \xi,\ u(\xi),\ \gamma)+u_1(\xi)\,v_1(x)\,u(\xi)]\,d\xi-sv_1(x) \qquad (35)$$

where s is an additional parameter. If we put

$$s=\int_a^b u_1(\xi)\,u(\xi)\,d\xi \qquad (36)$$

equation (35) becomes completely equivalent to (33). The result of this artificial transformation is that for $\gamma=\gamma_0$, $s=s_0=$ $=\int_a^b u_1(\xi)\,u_{\gamma_0}(\xi)\,d\xi$ and $u=u_{\gamma_0}(x)$ equation (35) satisfies the condition indicated in Sec. 7 under which the solution can be continued (with respect to two parameters γ and s).

Indeed, let the linearized kernel (analogous to (34)) corresponding to equation (35) have $\lambda=1$ as its characteristic value. This

means that the equation

$$u(x)=\int_a^b [f'_u(x,\ \xi,\ u_{\gamma_0}(\xi),\ \gamma_0)+u_1(\xi)\,v_1(x)]\,u(\xi)\,d\xi \tag{37}$$

has a nontrivial solution $U(x)$. Rewriting equation (37) for $u=U(x)$ in the form

$$u(x)=\int_a^b f'_u(x,\ \xi,\ u_{\gamma_0}(\xi),\ \gamma_0)\,u(\xi)\,d\xi+\int_a^b u_1(\xi)\,U(\xi)\,d\xi\cdot v_1(x) \tag{38}$$

we see that by Fredholm's third theorem this is only possible if the functions u_1 and U are orthogonal. But from (38) we then see that U must be proportional to u_1 and therefore these functions cannot be orthogonal to each other.

The contradiction thus obtained shows that for small $|\gamma-\gamma_0|$, $|s-s_0|$ the solution can be continued with respect to the parameters γ and s:

$$u=\psi(x,\ \gamma,\ s) \tag{39}$$

where $\psi(x,\ \gamma_0,\ s_0)=u_{\gamma_0}(x)$. Moreover, with the aid of the method of Sec. 7 this solution can be expanded into a series in powers of $(\gamma-\gamma_0)$ and $(s-s_0)$. Substituting solution (39) into the right-hand side of (36) we obtain the **bifurcation equation**

$$F(s,\ \gamma)\equiv s-\int_a^b u_1(\xi)\,\psi(\xi,\ \gamma,\ s)=0 \tag{40}$$

connecting s and γ. Finding $s(\gamma)$ from this equation and substituting the result into (39) we obtain a solution of equation (33).

However, differentiating equation (35) with respect to s and with respect to γ and reasoning as in the case of equation (38) we can readily prove (do it!) that

$$F'_s(s_0,\ \gamma_0)=0,$$

$$F'_\gamma(s_0,\ \gamma_0)=\frac{\int_a^b\int_a^b f'_\gamma(x,\ \xi,\ u_{\gamma_0}(\xi),\ \gamma_0)\,v_1(x)\,dx\,d\xi}{\int_a^b v_1^2(x)\,dx} \tag{41}$$

Hence, in the general case equation (40) cannot be uniquely resolved in s. If the double integral in (41) is different from zero equation (40) can be resolved with respect to γ, and then the relationship between s and γ is as shown in Fig. 119*a* or *b*. In the case (*a*) (which is the basic one) we have two solutions

for $\gamma < \gamma_0$ which, as γ increases, merge for $\gamma = \gamma_0$ and then, for $\gamma > \gamma_0$, become imaginary, that is disappear.

If the double integral in (41) is equal to zero we can follow the procedure described in Sec. II.2.8 in connection with equation (II.2.25). There is an important particular case here when, for all

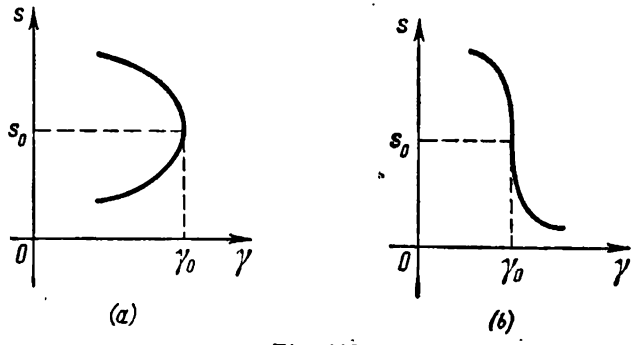


Fig. 119

values of γ, one of the branches $u_\gamma(x)$ of the solution of equation (33) is known, and we are interested in whether another branch can bifurcate from it for some value of γ. The corresponding value of γ determines a *bifurcation point* of the integral equation. Without loss of generality, we can assume that $u_\gamma(x) \equiv 0$. (We can always reduce the problem to this case by means of the substitution $u - u_\gamma(x) = v$.) Then equation (33) shows that

$$\int_a^b f(x, \xi, 0, \gamma)\, d\xi \equiv 0 \quad \text{and therefore} \quad \int_a^b f'_\gamma(x, \xi, 0, \gamma_0) d\xi = 0$$

Hence, the double integral in (41) is equal to zero, and therefore the expansion of the function $F(s, \gamma)$ in powers of s ($s_0 = 0$) and $(\gamma - \gamma_0)$ contains no linear terms. In the general case the quadratic terms are usually nonzero, and, according to (36), the branch $s(\gamma)$ in question is $s \equiv 0$; therefore, the shape of line (40) in the vicinity of the point $(\gamma_0, 0)$ shown in Fig. 120 is typical. This means that at the point $\gamma = \gamma_0$ there is a *bifurcation*, that is the solution can be extended in two ways, both when γ increases and decreases.

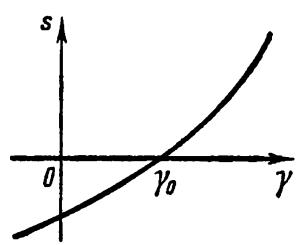


Fig. 120

There can also exist solutions whose shape differs from that in Fig. 119, however, it can be shown that if $\lambda = 1$ is a simple zero of Fredholm's determinant of the kernel $f'_u(x, \xi, 0, \gamma_0)$, then for $\gamma = \gamma_0$ there must be a bifurcation at

least on one side from γ_0. The same situation takes place for zeros of any odd multiplicity, but in the general case we must bear in mind that there may exist several nonzero branches. For zeros of even multiplicity there may be no bifurcation because the branches may be imaginary on both sides of γ_0. In some cases this multiplicity is inessential; for example, this is the case for equation (18) with positive definite symmetric kernel K if $\Phi(\xi, 0) \equiv 0$.

As an example, let us consider equation (4) having, in particular, a zero solution for all the values of the parameter λ. Since in this case the kernel $f'_u(x, \xi, 0, \lambda) = -2\lambda x$ has 1 as its characteristic value for $\lambda_0 = -1$, and $u_1(x) = x$, $v_1(x) = 1$ (check it up!), equation (35) takes the form

$$u(x) = \int_0^1 \left\{ -\frac{\lambda}{2} x [u^3(\xi) + 3u^2(\xi) + 4u(\xi)] + \xi u(\xi) \right\} d\xi - s \qquad (42)$$

Here it is convenient to put $\lambda + 1 = \mu$; then $\mu = \mu_0 = 0$ for $\lambda = \lambda_0 = -1$. Substituting the expansion

$$u(x) = a(x)\, s + b(x)\, \mu + c(x)\, s^2 + d(x)\, s\mu + e(x)\, \mu^2 + \ldots$$

into (42) we readily find (check it up!) that

$$a(x) = 3x, \quad b(x) = 0, \quad c(x) = -\frac{27}{8} x - \frac{9}{4}, \quad d(x) = \frac{9}{4} x + \frac{3}{2}, \quad e(x) = 0$$

whence the equation of bifurcation (40) takes the form $\frac{9}{4} s^2 - \frac{3}{2} \mu s + \ldots = 0$ where the dots designate the terms of degree not less than the third. Consequently, for $\mu = 0$, the law of bifurcation is $s = \frac{2}{3} \mu + O(\mu^2)$, that is $u(x) = 2x(\lambda + 1) + O((\lambda + 1)^2)$. This bifurcation was mentioned in Sec. 1; the principal term of the expansion can be easily derived from (5).

If to the characteristic value $\lambda = 1$ of kernel (34) there correspond k ($k > 1$) eigenfunctions we can reason as in the case $k = 1$. To this end, auxiliary equations analogous to (35) and (36) with additional parameters $s_1, s_2, \ldots, s_k$ are introduced, after which the corresponding system of k equations of bifurcation connecting the parameters $s_1, s_2, \ldots, s_k, \lambda$ is formed. In the general case the investigation of such systems is considerably more complicated than in the case of one equation with two parameters. Nevertheless, there are cases when successive elimination of unknown quantities leads to one equation, which makes it possible to complete the solution of the problem.

In the singular case under consideration we can directly apply to equation (33) the method of undetermined coefficients (in this case known as the **Nazarov-Nekrasov method**). As was shown in

Sec. II.2.8, when the theorem on the existence of an implicit function is inapplicable to a finite equation, one of the two quantities entering into the equation can be expressed in the form of a sum of a Puiseux series in fractional powers of the other quantity. Therefore, for small $|\gamma-\gamma_0|$, it is natural to seek the solution $s(\gamma)$ of the bifurcation equation and, together with it, the solution $u(x)$ of equation (33) in the form of the sum of a series:

$$u(x)=u_{\gamma_0}(x)+w_1(x)\,\delta+w_2(x)\,\delta^2+\ \ldots\ \ (\delta^p=\gamma-\gamma_0)$$

The integer $p\geqslant 1$ is not known in advance and is specified in the process of constructing the solution. The substitution of the series into (33) yields

$$u_{\gamma_0}(x)+w_1(x)\,\delta+w_2(x)\,\delta^2+\ldots=$$

$$=\int_a^b f(x,\ \xi,\ u_{\gamma_0}(\xi)+w_1(\xi)\,\delta+w_2(\xi)\,\delta^2+\ldots,\ \gamma_0+\delta^p)\,d\xi$$

Setting a value of p we expand the right-hand side in powers of δ. Then, equating the coefficients in like powers of δ, we obtain equations of the form

$$w_j(x)=\int_a^b f'_u(x,\ \xi,\ u_{\gamma_0}(\xi),\ \gamma_0)\,w_j(\xi)\,d\xi+\varphi_j(x) \qquad (43)$$

$$(j=1,\ 2,\ 3,\ \ldots)$$

where $\varphi_1(x)$ is a completely determined function while each of the subsequent functions $\varphi_j\ (j>1)$ is expressed in terms of the functions $w_1(x),\ w_2(x),\ \ldots,\ w_{j-1}(x)$.

Since for equation (43) the second Fredholm alternative takes place, the function $\varphi_1(x)$ must satisfy the corresponding k orthogonality conditions; if they are not satisfied, this means that p was chosen improperly. If they are satisfied, equation (43) with $j=1$ determines $w_1(x)$ as a function involving k arbitrary constants. Hence, the function $\varphi_2(x)$ also depends on k arbitrary constants, and for it to be possible to determine $w_2(x)$ from equation (43) with $j=2$, the function $\varphi_2(x)$ must satisfy the k orthogonality conditions corresponding to this equation from which these constants are found. Then $w_2(x)$ is determined to within k new arbitrary constants which enter into $\varphi_3(x)$ and, consequently, are determined from another k orthogonality conditions, etc. It can be shown that, beginning with a certain step, the equations determining the arbitrary constants become (algebraic) linear and have one and the same nondegenerate matrix, that is these systems possess uniquely determined solutions. After this step has been achieved inconsistent systems cannot occur, and therefore, at least in principle, it is possible to construct any number of terms of the expansion of the sought-for solution.

CHAPTER VIII

Ordinary Differential Equations

Fundamentals of the theory of ordinary differential equations are usually included into general courses of higher mathematics for engineering students. In the present chapter we shall consider some divisions of this theory which are not studied in these general courses but play an important role in modern theoretical and applied mathematics. The reader is supposed to be familiar with basic notions related to ordinary differential equations (e. g. see [107], Chapter XV).

§ 1. LINEAR EQUATIONS AND SYSTEMS

1. General Properties. In this section we consider linear systems of the form

$$\left.\begin{aligned} \dot{x}_1 &= a_{11}(t)\,x_1 + a_{12}(t)\,x_2 + \ldots + a_{1n}(t)\,x_n + f_1(t) \\ \dot{x}_2 &= a_{21}(t)\,x_1 + a_{22}(t)\,x_2 + \ldots + a_{2n}(t)\,x_n + f_2(t) \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \dot{x}_n &= a_{n1}(t)\,x_1 + a_{n2}(t)\,x_2 + \ldots + a_{nn}(t)\,x_n + f_n(t) \end{aligned}\right\} \qquad (1)$$

where the independent variable t is usually interpreted as time. Unless otherwise stated, we shall suppose that the independent variable is real while, in the general case, the coefficients and unknown functions may assume complex values. Introducing the vector-functions

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \\ \cdot \\ \cdot \\ \cdot \\ f_n \end{pmatrix}$$

and the coefficient matrix

$$\mathbf{A}=\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

we can rewrite system (1) in the form of the **vector differential equation**

$$\dot{\mathbf{x}}=\mathbf{A}(t)\,\mathbf{x}+\mathbf{f}(t) \tag{2}$$

As is known, linear differential equations and systems of linear equations of any order can always be reduced to systems of the first order, that is of form (2), by means of additional unknown functions.

We shall assume that the coefficients $a_{jk}(t)$ and the nonhomogeneity terms $f_j(t)$ are defined in an infinite or finite interval (which may be closed or nonclosed) of the variable t and take on finite values (by the way, we sometimes will allow them to turn into infinity at some points under the condition of their absolute integrability). Then for any initial condition

$$\mathbf{x}(t_0)=\mathbf{x}_0 \tag{3}$$

system (2) possesses exactly one solution defined over the whole interval in question.

Let us consider the homogeneous system

$$\dot{\mathbf{y}}=\mathbf{A}(t)\,\mathbf{y} \tag{4}$$

corresponding to system (2). The totality of all its solutions is an n-dimensional linear space. Every basis of this space, that is every collection

$$\mathbf{y}_1(t),\ \mathbf{y}_2(t),\ \dots,\ \mathbf{y}_n(t) \tag{5}$$

of n linearly independent solutions of system (4), is referred to as a **fundamental system of solutions.** It can readily be shown that the substitution of any fixed value t_0 into (5) yields a system of linearly independent number vectors. Indeed, if, for instance, $\mathbf{y}_1(t_0)=C_2\mathbf{y}_2(t_0)+\dots+C_n\mathbf{y}_n(t_0)$, then the two solutions $\mathbf{y}_1(t)$ and $C_2\mathbf{y}_2(t)+\dots+C_n\mathbf{y}_n(t)$ coincide at the point $t=t_0$ and therefore must coincide identically for all t (why?), which contradicts the hypothesis that (5) is a fundamental system of solutions.

We shall also use the notion of the linear homogeneous system **adjoint** to (4) which is written as

$$\dot{\mathbf{z}}=-\mathbf{A}^*(t)\,\mathbf{z} \tag{6}$$

It can easily be shown that for any solution $\mathbf{y}(t)$ of system (4) and any solution $\mathbf{z}(t)$ of system (6) the relation

$$(\mathbf{y},\ \mathbf{z})\equiv\mathbf{z}^*\mathbf{y}=\text{const}$$

holds. Virtually,

$$\frac{d}{dt}(\mathbf{y}, \mathbf{z}) = (\dot{\mathbf{y}}, \mathbf{z}) + (\mathbf{y}, \dot{\mathbf{z}}) = (\mathbf{Ay}, \mathbf{z}) + (\mathbf{y}, -\mathbf{Az}^*) \equiv 0$$

which is what we set out to prove. (Let the reader obtain this result by differentiating the expression $\mathbf{z}^*\mathbf{y}$.) If $\mathbf{A}(t) \equiv -\mathbf{A}^*(t)$, i.e. $\mathbf{A}$ is an anti-Hermitian matrix, system (4) is said to be **self-adjoint**.

Now consider the **matrix differential equation**

$$\dot{\mathbf{Y}} = \mathbf{A}(t)\,\mathbf{Y} \tag{7}$$

where $\mathbf{Y} = \mathbf{Y}(t)$ is a sought-for square matrix of order n. We can write this system in scalar form, which results in a system of n^2 scalar linear differential equations for the elements of the matrix $\mathbf{Y}$. It is readily seen that for the elements of every column of the matrix $\mathbf{Y}$ we thus obtain a system of n equations which can be written in vector form (4) (let the reader verify this fact for the first column of the matrix $\mathbf{Y}$). Thus, the general solution of equation (7) consists of all the matrices whose columns are n arbitrary vector functions satisfying equation (4). Now suppose that $\mathbf{Y}(t)$ is a solution of equation (7). Then the vector $\mathbf{Y}(t)\mathbf{c}$ where $\mathbf{c}$ is an arbitrary constant number vector is a solution of equation (4). Indeed, we have

$$\frac{d}{dt}(\mathbf{Yc}) = \dot{\mathbf{Y}}\mathbf{c} = \mathbf{A}(t)\,\mathbf{Yc} = \mathbf{A}(t)(\mathbf{Yc})$$

which is what we set out to prove (let the reader find out where the constancy of the vector $\mathbf{c}$ has been used in this proof).

The determinant of the matrix $\mathbf{Y}(t)$ is called the **Wronskian** (after J.M. Wronski (1778-1853), a Polish mathematician and philosopher, who introduced such determinants in 1812) of the corresponding collection (5) of solutions of the system of equations (4). Let us derive a simple formula for this determinant. We shall use the following formula for differentiation of a determinant whose elements are functions of an independent variable:

$$\begin{vmatrix} y_{11} & y_{12} & \dots & y_{1n} \\ y_{21} & y_{22} & \dots & y_{2n} \\ \dots & \dots & \dots & \dots \\ y_{n1} & y_{n2} & \dots & y_{nn} \end{vmatrix}' = \begin{vmatrix} y'_{11} & y'_{12} & \dots & y'_{1n} \\ y_{21} & y_{22} & \dots & y_{2n} \\ \dots & \dots & \dots & \dots \\ y_{n1} & y_{n2} & \dots & y_{nn} \end{vmatrix} + \begin{vmatrix} y_{11} & y_{12} & \dots & y_{1n} \\ y'_{21} & y'_{22} & \dots & y'_{2n} \\ y_{31} & y_{32} & \dots & y_{3n} \\ \dots & \dots & \dots & \dots \\ y_{n1} & y_{n2} & \dots & y_{nn} \end{vmatrix} + \dots + \begin{vmatrix} y_{11} & y_{12} & \dots & y_{1n} \\ y_{21} & y_{22} & \dots & y_{2n} \\ \dots & \dots & \dots & \dots \\ y_{n-1,1} & y_{n-1,2} & \dots & y_{n-1,n} \\ y'_{n1} & y'_{n2} & \dots & y'_{nn} \end{vmatrix}. \tag{8}$$

To prove this formula one must write explicitly the determinant on the left-hand side of (8), differentiate it termwise and then again group the summands and form the corresponding determinants (let the reader carry out this procedure for $n=2$ and $n=3$). Now suppose that each column on the left-hand side of (8) satisfies system (4). For simplicity, let us take the case $n=3$. Then the first summand on the right-hand side of (8) can be written in the form

$$\begin{vmatrix} a_{11}y_{11}+a_{12}y_{21}+a_{13}y_{31} & a_{11}y_{12}+a_{12}y_{22}+a_{13}y_{32} & a_{11}y_{13}+a_{12}y_{22}+a_{13}y_{33} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & y_{33} \end{vmatrix}$$

Multiplying the second row of this determinant by $-a_{12}$, the third by $-a_{13}$ and adding them to the first one we obviously obtain the quantity $a_{11}\det \mathbf{Y}(t)$. Transforming in like manner all the other terms on the right-hand side of (8) we obtain

$$\frac{d}{dt}[\det \mathbf{Y}(t)] = \operatorname{tr}\mathbf{A}(t)\ \det \mathbf{Y}(t)$$

whence

$$\det \mathbf{Y}(t) = \det \mathbf{Y}(t_0)\ \exp \int_{t_0}^{t} \operatorname{tr}\mathbf{A}(\tau)\, d\tau \tag{9}$$

where $\operatorname{tr}\mathbf{A}$ is the **trace** of the matrix $\mathbf{A}$ determined by the formula

$$\operatorname{tr}\mathbf{A} = a_{11}+a_{22}+\ldots+a_{nn}$$

This quantity is also denoted as $\operatorname{Sp}\mathbf{A}$ (German "Spur" trace).

Formula (9) was obtained in 1827 by N.H. Abel for second-order equations and in 1838 by J. Liouville and M.V. Ostrogradsky for the general case.

If a solution $\mathbf{Y}(t)$ of equation (7) satisfies the condition $\det \mathbf{Y}(t) \not\equiv 0$ it is called a **fundamental matrix** of system (4). Apparently, $\mathbf{Y}(t)$ is a fundamental matrix if and only if its columns form a fundamental system of solutions of equation (4). For practical purposes the so-called **principal matrix solution** of system (4) which is a fundamental matrix satisfying the condition

$$\mathbf{Y}(t_0) = \mathbf{I}$$

is particularly convenient. We shall denote the principal matrix solution by the symbol $\mathbf{Y}(t, t_0)$ (and thus $\mathbf{Y}(t_0, t_0) \equiv \mathbf{I}$). If $\mathbf{y}(t)$ is a solution of system (4) satisfying an initial condition $\mathbf{y}(t_0) = \mathbf{y}_0$ it is expressed in terms of $\mathbf{Y}(t, t_0)$ by means of the formula

$$\mathbf{y}(t) = \mathbf{Y}(t, t_0)\,\mathbf{y}_0 \tag{10}$$

(let the reader prove this result). The matrix $\mathbf{Y}(t, t_0)$ satisfies the relation

$$\mathbf{Y}(t, t_0) \equiv \mathbf{Y}(t, t_1)\,\mathbf{Y}(t_1, t_0) \tag{11}$$

for any t, t_0 and t_1. In fact, both members of this equality satisfy equation (7) and, besides, they coincide at the point $t = t_1$ (why?), which implies the validity of (10). Putting $t = t_0$ we obtain another important formula:

$$\mathbf{Y}(t_1, t_0) = [\mathbf{Y}(t_0, t_1)]^{-1}$$

For an arbitrary constant vector $\mathbf{c}$ the expression $\mathbf{Y}(t, t_0)\mathbf{c}$ is the general solution of system (4) because it satisfies this system and involves n arbitrary parameters (the components of $\mathbf{c}$). Therefore, according to the well-known method of variation of arbitrary constants (e.g. see [107]. Sec. XV.21), the solution of nonhomogeneous system (2) should be looked for in the form $\mathbf{x} = \mathbf{Y}(t, t_0)\boldsymbol{\varphi}(t)$ where $\boldsymbol{\varphi}(t)$ is a vector function of t to be determined. Substituting this expression into (2) we obtain

$$\frac{d}{dt}(\mathbf{Y}\boldsymbol{\varphi}) = \dot{\mathbf{Y}}\boldsymbol{\varphi} + \mathbf{Y}\dot{\boldsymbol{\varphi}} = \mathbf{A}\mathbf{Y}\boldsymbol{\varphi} + \mathbf{Y}\dot{\boldsymbol{\varphi}} = \mathbf{A}\mathbf{Y}\boldsymbol{\varphi} + \mathbf{f}$$

whence

$$\dot{\boldsymbol{\varphi}} = \mathbf{Y}^{-1}\mathbf{f}, \quad \boldsymbol{\varphi}(t) = \int_{t_0}^{t} [\mathbf{Y}(\tau, t_0)]^{-1}\mathbf{f}(\tau)\,d\tau + \mathbf{c}$$

Consequently, by virtue of (11), we have

$$\mathbf{x}(t) = \int_{t_0}^{t} \mathbf{Y}(t, \tau)\mathbf{f}(\tau)\,d\tau + \mathbf{Y}(t, t_0)\mathbf{c} \tag{12}$$

The substitution of $t = t_0$ into the right-hand side of (12) yields $\mathbf{c}$, and therefore, to obtain the solution satisfying initial condition (3) we must simply replace $\mathbf{c}$ by $\mathbf{x}_0$ in formula (12).

The matrix $\mathbf{Y}(t, t_0)$ can be written as a sum of a series. To this end, we substitute $\mathbf{Y}(t, t_0)$ into (7) and integrate both members from t_0 to t. Since $\mathbf{Y}(t_0, t_0) = \mathbf{I}$ this results in the integral equation

$$\mathbf{Y}(t, t_0) = \mathbf{I} + \int_{t_0}^{t} \mathbf{A}(\tau)\,\mathbf{Y}(\tau, t_0)\,d\tau$$

Now, iterating this relation we arrive at the formula (check it up!)

$$\mathbf{Y}(t, t_0) = \mathbf{I} + \int_{t_0}^{t} \mathbf{A}(\tau_1)\,d\tau_1 + \int_{t_0}^{t} \mathbf{A}(\tau_1)\,d\tau_1 \int_{t_0}^{\tau_1} \mathbf{A}(\tau_2)\,d\tau_2 + \\ + \int_{t_0}^{t} \mathbf{A}(\tau_1)\,d\tau_1 \int_{t_0}^{\tau_1} \mathbf{A}(\tau_2)\,d\tau_2 \int_{t_0}^{\tau_2} \mathbf{A}(\tau_3)\,d\tau_3 + \ldots \tag{13}$$

Let the reader show that the principal matrix solution of the adjoint system (6) is expressed by the formula

$$\mathbf{Z}(t, t_0) = [\mathbf{Y}(t_0, t)]^* \tag{14}$$

The results presented here become particularly simple when the coefficient matrix $\mathbf{A}$ is constant. In this case we have $\mathbf{Y}(t, t_0) = e^{(t-t_0)\mathbf{A}}$ (e.g. see [107], Sec. XVII.18), and formula (12) takes the form

$$\mathbf{x}(t) = \int_{t_0}^{t} e^{(t-\tau)\mathbf{A}} \mathbf{f}(\tau)\, d\tau + e^{(t-t_0)\mathbf{A}} \mathbf{c}$$

(let the reader verify this formula directly by substituting the expression of $\mathbf{x}(t)$ into (2) for $\mathbf{A} = \text{const}$). The form of the fundamental matrix and the rules for computing functions of matrices (see Sec. IV.3.5) imply the following proposition: *for all the solutions of the equation* $\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}$ *with constant matrix* $\mathbf{A}$ *to tend to zero for* $t \to \infty$ *it is necessary and sufficient that all the eigenvalues* λ_j $(j = 1, 2, \ldots, n)$ *of the matrix* $\mathbf{A}$ *have negative real parts. For all the solutions to be bounded as* $t \to \infty$ *it is necessary and sufficient that* $\operatorname{Re}\lambda_j \leqslant 0$ $(j = 1, 2, \ldots, n)$ *and that the Jordan submatrices corresponding to the eigenvalues with zero real parts be of the first order* (see formula (IV.3.13)).

2. Periodic Systems. Now let us suppose that the coefficients of system (4) are periodic functions of t with one and the same period $T > 0$, that is

$$\mathbf{A}(t+T) \equiv \mathbf{A}(t)$$

Then we have

$$\mathbf{Y}(t+T, t_0+T) \equiv \mathbf{Y}(t, t_0) \tag{15}$$

because if the initial time moment entering into the initial conditions is shifted by a period the solution undergoes the same transformation. Our next task is to prove that there exists a constant matrix $\mathbf{M}$ such that the matrix function $\mathbf{Y}(t, 0)e^{-t\mathbf{M}}$ is periodic in t with period T. Note that if $\mathbf{Y}(t, 0)e^{-t\mathbf{M}} = \mathbf{R}(t)$ where $\mathbf{R}(t)$ is periodic with period T we obtain the following result: *the principal matrix solution of periodic system* (4) *is representable in the form*

$$\mathbf{Y}(t, 0) = \mathbf{R}(t)\, e^{t\mathbf{M}} \tag{16}$$

where the matrix function $\mathbf{R}(t)$ *is periodic with period* T (this proposition is known as **Floquet's theorem**).

To prove the above assertion about the existence of the matrix $\mathbf{M}$ we use properties (11) and (15) according to which the required condition that the matrix $\mathbf{Y}(t, 0)e^{-t\mathbf{M}}$ should be periodic can be rewritten as

$$\mathbf{Y}(t, 0)e^{-t\mathbf{M}} = \mathbf{Y}(t+T, 0)e^{-(t+T)\mathbf{M}} =$$
$$= \mathbf{Y}(t+T, T)\,\mathbf{Y}(T, 0)e^{-T\mathbf{M}}e^{-t\mathbf{M}} = \mathbf{Y}(t, 0)\,\mathbf{Y}(T, 0)e^{-T\mathbf{M}}e^{-t\mathbf{M}}$$

which yields

$$\mathbf{Y}(T, 0) = e^{T\mathbf{M}}, \text{ that is } \mathbf{M} = \frac{1}{T}\ln \mathbf{Y}(T, 0)$$

where the logarithm of the matrix $\mathbf{Y}(T, 0)$ is understood in accordance with the rule given in Sec. IV.3.5. The matrix $\mathbf{M}$ being constructed, Floquet's theorem have thus been proved.

Formula (16) indicates that *for all the solutions of the periodic system* (4) *to tend to zero for* $t \to \infty$ *it is necessary and sufficient that all the eigenvalues of the matrix* $\mathbf{M}$ *have negative real parts, that is all the eigenvalues of the matrix* $\mathbf{Y}(T, 0)$ *(the* ***monodromy matrix****) be less than unity in their moduli.* The necessary and sufficient conditions for the solutions of periodic system (4) to be bounded for $t \to \infty$ are formulated by analogy with the end of Sec. 1. The eigenvalues of the monodromy matrix $\mathbf{Y}(T, 0)$ are called **(characteristic) multipliers** of the periodic system (4). We have

$$\mathbf{Y}(T+t_0, t_0) = \mathbf{Y}(T+t_0, T)\,\mathbf{Y}(T, 0)\,\mathbf{Y}(0, t_0) = \\ = \mathbf{Y}(t_0, 0)\,\mathbf{Y}(T, 0)\,[\mathbf{Y}(t_0, 0)]^{-1}$$

and therefore the eigenvalues and the eigenvectors of the monodromy matrix are independent of the choice of the initial time moment (why?).

With every multiplier ρ and the corresponding eigenvector $\mathbf{d}$ we can associate the solution $\mathbf{y} = \mathbf{Y}(t, 0)\,\mathbf{d}$ which possesses the following property:

$$\mathbf{y}(t+T) = \mathbf{Y}(t+T, 0)\,\mathbf{d} = \mathbf{Y}(t+T, T)\,\mathbf{Y}(T, 0)\,\mathbf{d} = \\ = \mathbf{Y}(t, 0)\,\rho\mathbf{d} = \rho\mathbf{y}(t) \qquad (17)$$

This accounts for the term "multiplier". For every concrete periodic system the multipliers can be computed by means of numerical integration of the system. If the coefficients of the system are close to constants it is expedient to use the small-parameter method (let the reader apply the small-parameter scheme to the system $\dot{\mathbf{y}} = (\mathbf{A} + \alpha\mathbf{B}(t))\,\mathbf{y}$ where $|\alpha|$ is small, $\mathbf{A}$ is constant and $\mathbf{B}(t)$ is periodic).

The product of all the eigenvalues of a matrix is equal to its determinant (why?), and therefore (9) implies that the product of all multipliers (each multiplier is counted according to its multiplicity) is expressed by the formula

$$\prod_{j=1}^{n} \rho_j = \exp \int_0^T \operatorname{tr} \mathbf{A}(t)\,dt \qquad (18)$$

Consequently, if $\operatorname{Re} \int_0^T \operatorname{tr} \mathbf{A}(t)\,dt > 0$, there is at least one multiplier ρ_j whose modulus exceeds unity, and then the corresponding solution of form (17) is unbounded for $t \to \infty$. (It can also be proved that if at least one solution is unbounded then all solu-

tions, except possibly a family of solutions with less than n degrees of freedom, are unbounded as well; this property follows from the form of the general solution of a linear system.) Similarly, if $\operatorname{Re}\int_0^T \operatorname{tr} \mathbf{A}(t)\,dt < 0$, system (4) possesses solutions which are unbounded for $t \to -\infty$. It follows that for all the solutions of periodic system (4) to be bounded on the entire t-axis it is necessary (but not sufficient!) that $\operatorname{Re}\int_0^T \operatorname{tr} \mathbf{A}(t)\,dt = 0$.

The last condition is, in particular, fulfilled in the case of the system

$$\dot{y}_1 = y_2,\ \dot{y}_2 = -p(t)\,y_1 \qquad (\text{i.e. } \ddot{y}_1 + p(t)\,y_1 = 0) \tag{19}$$

with periodic coefficient $p(t)$: $p(t+T) \equiv p(t)$. This system was dealt with in many theoretical and applied articles. It possesses two multipliers whose product, according to formula (18), is equal to unity. Series (13) for the matrix $\mathbf{A}(t) = \begin{pmatrix} 0 & 1 \\ -p(t) & 0 \end{pmatrix}$ was analysed by A.M. Lyapunov who proved that if the function $p(t)$ is nonnegative and satisfies the condition

$$0 < \int_0^T p(t)\,dt \leqslant \frac{4}{T}$$

the characteristic equation

$$\rho^2 - (y_{11}(T,\ 0) + y_{22}(T,\ 0))\,\rho + 1 = 0 \tag{20}$$

of the monodromy matrix has a negative discriminant, and therefore $|\rho_1| = |\rho_2| = 1$ and $\rho_1 \neq \rho_2$, which means that every solution of system (19) is bounded on the whole t-axis.

Let us make in periodic system (4) the substitution $\mathbf{y} = \mathbf{R}(t)\,\mathbf{z}$ where $\mathbf{R}(t)$ is the matrix entering into formula (16) and $\mathbf{z}$ is the new unknown vector function; then the resultant equation for $\mathbf{z}$ is of the form $\dot{\mathbf{z}} = \mathbf{M}\mathbf{z}$ (check it up!). Hence, system (4) is reducible to a linear system with constant coefficients. Generally, if we have a system of type (4) (not necessarily a periodic one) and there exists a linear change of variables of the form $\mathbf{y} = \mathbf{H}(t)\,\mathbf{z}$ (where the matrix function $\mathbf{H}(t)$ is nondegenerate and bounded for all t together with its inverse matrix $[\mathbf{H}(t)]^{-1}$) which transforms system (4) to a system with constant coefficients we say that system (4) is **reducible.** Given a reducible system, it is possible to judge upon the properties of its solutions (in particular, upon their boundedness and their behaviour for $t \to \pm\infty$) on the basis of the properties of the solutions of the corresponding system

with constant coefficients. The results of this section thus mean that *every periodic linear system is reducible.*

There are many interesting results concerning the problem of reducibility of periodic and even almost periodic (Sec. 7) systems whose coefficients are close to constants. For such systems the transformation matrix $\mathbf{H}(t)$ is constructed in the form of an asymptotically convergent series (see Sec. II.4.1) in powers of a small parameter and the reduction is also understood in the asymptotic sense.

On reducible systems we refer the reader to [35] and [119].

3. Hill's Equation. The equation

$$\ddot{y}+[\lambda+\mu f_1(t)+\mu^2 f_2(t)+\ldots]\,y=0 \tag{21}$$

in which all the functions $f_j(t)$ are periodic with the same period is called **Hill's equation**. For the sake of simplicity, we shall suppose that $f_j(t+\pi)\equiv f_j(t)$, $j=1, 2, \ldots$ (this assumption does not restrict the generality because we can always come to this case by means of a transformation of the form $t\longrightarrow kt$). All the quantities involved will be regarded as real. If the expression in the square brackets is of the form $\lambda+\mu\cos 2t$ we obtain the important particular case known as the **Mathieu equation**. In this section we shall discuss the question of boundedness of the solutions of equation (21) on the whole t-axis depending on the values of the parameters λ and μ.

Let $\varphi(t, \lambda, \mu)$ and $\psi(t, \lambda, \mu)$ be the solutions of equation (21) satisfying, respectively, the initial conditions $y(0)=1$, $\dot{y}(0)=0$ and $y(0)=0$, $\dot{y}(0)=1$. Using the notation of Sec. 1 we can write $\varphi=y_{11}(t, 0)$ and $\psi=y_{12}(t, 0)$. Besides, let us introduce the quantity

$$A=A(\lambda, \mu)=\frac{1}{2}[\varphi(\pi, \lambda, \mu)+\dot{\psi}(\pi, \lambda, \mu)] \tag{22}$$

The characteristic equation (20) shows that in this case we have $|\rho_1|=|\rho_2|=1$ and $\rho_1\neq\rho_2$ for $|A|<1$, which means that every solution of equation (21) is bounded on the whole t-axis. Accordingly, if $A>1$ ($A<-1$), then $0<\rho_1<1<\rho_2$ ($\rho_2<-1<\rho_1<0$), that is every nonzero solution of equation (21) is unbounded. Thus, the problem reduces to expressing these results in terms of the given functions $f_j(t)$ and the values of the parameters λ and μ.

Let us suppose that $|\mu|$ is small and expand the function φ in powers of μ:

$$\varphi(t, \lambda, \mu)=\varphi_0(t, \lambda)+\varphi_1(t, \lambda)\,\mu+\varphi_2(t, \lambda)\,\mu^2+\ldots$$

Substituting this expansion into equation (21) and into the corresponding initial conditions and equating the coefficients in like

powers of μ we obtain the relations

$$\begin{aligned}
&\ddot{\varphi}_0+\lambda\varphi_0=0, && \varphi_0|_{t=0}=1, \quad \dot{\varphi}_0|_{t=0}=0\\
&\ddot{\varphi}_1+\lambda\varphi_1=-\varphi_0 f_1, && \varphi_1|_{t=0}=0, \quad \dot{\varphi}_1|_{t=0}=0\\
&\ddot{\varphi}_2+\lambda\varphi_2=-\varphi_0 f_2-\varphi_1 f_1, && \varphi_2|_{t=0}=0, \quad \dot{\varphi}_2|_{t=0}=0\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots
\end{aligned}$$

which enable us to determine, in succession, the functions φ_0, φ_1, φ_2, etc. (it is readily seen that $\varphi_0=\cos\sqrt{\lambda}t$). Analogously, expanding the function ψ we find that $\psi_0=\frac{1}{\sqrt{\lambda}}\sin\sqrt{\lambda}t$ and that the subsequent recurrence equations for ψ_j ($j=1, 2, \ldots$) are the same as for the functions φ_j ($j=1, 2, \ldots$). Hence, by virtue of (22), we obtain

$$A=\cos\sqrt{\lambda}\pi+\frac{1}{2}[\varphi_1(\pi, \lambda)+\dot{\psi}_1(\pi, \lambda)]\mu+\ldots \qquad (23)$$

According to the above, the regions in the $\lambda\mu$-plane for which the solutions of equation (21) are bounded are separated from the regions of unbounded solutions by a critical line whose equation is $[A(\lambda, \mu)]^2=1$. Therefore, formula (23) implies that this line intersects the λ-axis at the points $\lambda=n^2$ ($n=0, 1, 2, \ldots$), the parts of the λ-axis between these points belonging to the region of boundedness. It can be shown (but we do not give the proof here) that for $n>0$ these points are the points of self-intersection of the critical line (see Fig. 121 where the regions of unboundedness of the solutions are shaded).

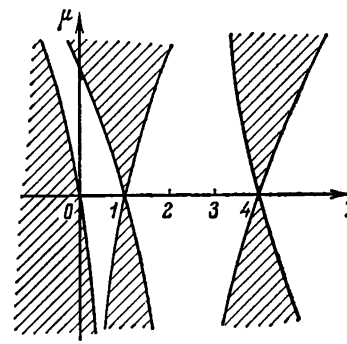


Fig. 121

To construct these regions we can apply the following technique. Note that for the branches of the critical line passing through the points $\lambda=n^2$ with even values of n we have $A=1$, and therefore the corresponding multipliers are $\rho_1=\rho_2=1$. Consequently, for each point of these branches it is possible to construct a periodic solution $y(t, \mu)$ of equation (21) with period π which is independent of λ since on these branches λ is a function of μ. Following the idea of the method of undetermined coefficients we now take the expansions

$$y=y_0(t)+y_1(t)\mu+y_2(t)\mu^2+\ldots, \quad \lambda=n^2+a_1\mu+a_2\mu^2+\ldots \qquad (24)$$

for $n=0, 2, 4, \ldots$ (for each n the corresponding branch of the critical line passes through the point $(n^2, 0)$), substitute them into

(21) and equate the coefficients in the same powers of μ, which results in the equations

$$\ddot{y}_0 + n^2 y_0 = 0,$$

$$\ddot{y}_1 + n^2 y_1 = -[a_1 + f_1(t)]\, y_0(t), \tag{25}$$

$$\ddot{y}_2 + n^2 y_2 = -[a_1 + f_1(t)]\, y_1(t) - [a_2 + f_2(t)]\, y_0(t) \tag{26}$$

. . . . . .

From the first equation we obtain

$$y_0 = A \cos nt + B \sin nt$$

where A and B are constant parameters. The right member of (25) being periodic with period π, we can expand it in Fourier's series with respect to the functions $\cos 2kt$ and $\sin 2kt$. Since the function $y_1(t)$ must also be π-periodic, the corresponding condition guaranteeing the absence of resonance should hold. This means that the Fourier expansion of the right member of (25) must not contain the terms with $\cos nt$ and $\sin nt$, which yields the equalities

$$\left.\begin{aligned} &\int_0^{\pi} [a_1 + f_1(t)]\,(A \cos nt + B \sin nt) \cos nt\, dt = 0, \\ &\int_0^{\pi} [a_1 + f_1(t)]\,(A \cos nt + B \sin nt) \sin nt\, dt = 0 \end{aligned}\right\} \tag{27}$$

Taking the values $n = 2, 4, 6, \ldots$ we derive from (27) the equations

$$\left.\begin{aligned} (\alpha + a_1) A + \beta B &= 0, \\ \beta A + (\gamma + a_1) B &= 0 \end{aligned}\right\} \tag{28}$$

where

$$\alpha = \frac{2}{\pi} \int_0^{\pi} f_1(t) \cos^2 nt\, dt, \quad \beta = \frac{2}{\pi} \int_0^{\pi} f_1(t) \cos nt \sin nt\, dt$$

and

$$\gamma = \frac{2}{\pi} \int_0^{\pi} f_1(t) \sin^2 nt\, dt$$

For system (28) to possess a nonzero solution it is necessary and sufficient that

$$\begin{vmatrix} \alpha + a_1 & \beta \\ \beta & \gamma + a_1 \end{vmatrix} = 0 \tag{29}$$

whence we determine two real (why?) values of a_1 corresponding to the two directions along which the critical line intersects the

λ-axis at the point $\lambda = n^2$. According to (28), to each of these values there correspond some values of A and B determined to within a proportionality factor. But since the solution $y(t, \mu)$ itself is defined to within a scalar factor, we can introduce for y a normalization condition. For example, we can put $y(0, \mu) = 1$, which implies

$$y_0(0) = 1, \ y_1(0) = 0, \ y_2(0) = 0, \ \ldots \tag{30}$$

The first of these equalities determines the coefficients A and B uniquely.

Conditions (27) thus being satisfied, we can write the general solution of equation (25) in the form

$$y_1 = Y_1(t) + A_1 \cos nt + B_1 \sin nt$$

where $Y_1(t)$ is a periodic function of period π and A_1, B_1 are arbitrary constants. These arbitrary constants and the coefficient a_2 are then found from the nonresonance condition for equation (26) and from the second equality (30), and so on. Proceeding in this way we can determine any number of terms in expansion (24). (It should be noted that there are cases when the roots of equation (29) coincide; then to distinguish between the branches of the critical line intersecting the λ-axis we should resort to the terms with higher powers of μ. In particular, for $n \geqslant 2$ this is the case in the Mathieu equation). Generally speaking, this construction yields only one periodic solution of equation (21) while the other linearly independent solution is usually of the form $ty(t, \mu) + \tilde{y}(t, \mu)$ (where $\tilde{y}(t + 2\pi, \mu) \equiv \tilde{y}(t, \mu)$), i.e. unbounded.

For the points in $\lambda\mu$-plane to which there correspond unbounded solutions (these regions are shaded in Fig. 121) we can readily derive asymptotic expansions of the multipliers in powers of μ. According to Sec. 2, the unbounded solutions are of the form

$$y = e^{\pm st} R(t) \qquad (e^{\pm s\pi} = \rho_{1,2}; \ R(t + \pi) \equiv R(t)) \tag{31}$$

where s is real (why?). We have

$$s|_{\mu=0} = 0, \ R(t)|_{\mu=0} = A \cos nt + B \sin nt$$

and therefore, when substituting (31) into (21), we put $\lambda = n^2 + \lambda_1\mu + \lambda_2\mu^2 + \ldots$ and

$$s = s_1\mu + s_2\mu^2 + \ldots, \ R(t) = A \cos nt + B \sin nt + R_1(t)\mu + R_2(t)\mu^2 + \ldots \tag{32}$$

Equating the coefficients in like powers of μ after the substitution has been performed we obtain

$$\ddot{R}_1 + n^2 R_1 = -[\lambda_1 + f_1(t)](A \cos nt + B \sin nt) \mp \mp 2ns_1(A \sin nt + B \cos nt)$$

whence, by the condition that there must be no resonance, we find (check it up!) that

$$\left.\begin{aligned}(\lambda_1+\alpha)\,A+(\mp 2ns_1+\beta)\,B=0\\(\pm 2ns_1+\beta)\,A+(\lambda_1+\gamma)\,B=0\end{aligned}\right\} \tag{33}$$

where α, β and γ are understood in the sense of (28). From (33) we obtain

$$\begin{vmatrix}\lambda_1+\alpha & \mp 2ns_1+\beta\\ \pm 2ns_1+\beta & \lambda_1+\gamma\end{vmatrix}=0 \tag{34}$$

that is

$$4n^2s_1^2-\beta^2+(\lambda_1+\alpha)(\lambda_1+\gamma)=0$$

It follows that

$$s_1=\frac{1}{2n}\sqrt{\beta^2-(\lambda_1+\alpha)(\lambda_1+\gamma)}$$

For the right member of this equality to be real it is necessary that λ_1 be contained in the interval

$$-\frac{\alpha+\gamma}{2}-\sqrt{\frac{(\alpha+\gamma)^2}{4}+\beta^2}\leqslant\lambda_1\leqslant-\frac{\alpha+\gamma}{2}+\sqrt{\frac{(\alpha+\gamma)^2}{4}+\beta^2}$$

Now, from (33) we can determine the ratio $A:B$. The subsequent terms of expansion (32) are obtained in a similar way. These expansions can be found in [55].

In the case $n=0$ we have $y_0\equiv 1$, and therefore, instead of (27), we obtain the equality $\int\limits_0^{\pi}[a_1+f_1(t)]\,dt=0$ which yields only one value

$$a_1=-\frac{1}{\pi}\int\limits_0^{\pi}f_1(t)\,dt$$

The further calculations remain similar to the above.

If the number n is odd we have $A=-1$ and $\rho_{1,2}=-1$. In this case the solution of equation (21) satisfies the relation $y(t+\pi)\equiv\equiv -y(t)$ whence $y(t+2\pi)\equiv y(t)$. Consequently, we can again write down expansion (24) and perform the same calculations under the assumption that all the functions involved are periodic with period 2π.

For example, let us consider the critical line for the Mathieu equation in the vicinity of the point $\lambda=1^2$. We obtain the relation

$$\begin{aligned}\ddot{y}_0+\ddot{y}_1\mu+\ddot{y}_2\mu^2+\ldots+(1+a_1\mu+a_2\mu^2+\ldots+\\+\mu\cos 2t)(y_0+y_1\mu+y_2\mu^2+\ldots)=0\end{aligned} \tag{35}$$

from which, as above, it follows that

$$\ddot{y}_0 + y_0 = 0, \text{ that is } y_0 = A\cos t + B\sin t$$
$$\ddot{y}_1 + y_1 = -(a_1 + \cos 2t)(A\cos t + B\sin t) \tag{36}$$

etc. The condition of the absence of resonance then yields (check it up!)

$$\left(a_1 + \frac{1}{2}\right)A = 0, \qquad \left(a_1 - \frac{1}{2}\right)B = 0$$

whence $a_1 = \pm\frac{1}{2}$. Now note that under the transformation $\mu \longrightarrow -\mu$, $t \longrightarrow \frac{\pi}{2} + t$ the Mathieu equation goes into itself, and therefore it is sufficient to construct only one branch of the critical line because the other is obtained from that branch by means of the transformation $\mu \longrightarrow -\mu$. Let us put $a_1 = -\frac{1}{2}$; then we obtain $B = 0$ and, by virtue of (30), $A = 1$ (if we take $a_1 = \frac{1}{2}$ the normalization condition for y should be changed; for instance, it can be taken in the form $\dot{y}(0, \mu) = 1$). From (36) we find

$$y_1 = \frac{1}{16}\cos 3t + A_1\cos t + B_1\sin t$$

Now, equating in (35) the coefficients in μ^2 we obtain the equation

$$\ddot{y}_2 + y_2 = -(a_1 + \cos 2t)\, y_1(t) - a_2 y_0(t) =$$
$$= -\left(-\frac{1}{2} + \cos 2t\right)\left(\frac{1}{16}\cos 3t + A_1\cos t + B_1\sin t\right) - a_2\cos t$$

The condition that there are no resonant terms in $y_2(t)$ yields $a_2 = -\frac{1}{32}$ and $B_1 = 0$ (check it up!) while the second equality (30) results in $A_1 = -\frac{1}{16}$. Confining ourselves to these computations we can write the equation of the critical line in the form

$$\lambda = 1 - \frac{1}{2}\mu - \frac{1}{32}\mu^2 + \ldots$$

To each branch of the critical line for the Mathieu equation passing through the point $\lambda = n^2$, $\mu = 0$ there corresponds exactly one (to within a constant factor) even or odd periodic solution turning, respectively, into $\cos nt$ or $\sin nt$ for $\mu = 0$. Under the normalization condition given below these solutions are denoted, respectively, as $\mathrm{ce}_n(t, q)$ and $\mathrm{se}_n(t, q)$ (where $q = 16\mu$) and called the **Mathieu functions.** The normalization condition requires that the coefficient in $\cos nt$ in the Fourier series of $\mathrm{ce}_n(t, q)$ and the coefficient in $\sin nt$ in the Fourier series for $\mathrm{se}_n(t, q)$ should be

equal to unity. For the theory and applications of the Mathieu functions we refer the reader to [96].

4. Parametric Resonance. For sufficiently small values of $|\mu|$ the coefficient $\lambda+\mu f_1(t)+\ldots$ in equation (21) can be made as close as desired to the constant λ, and if this coefficient is exactly equal to a constant number all the solutions are bounded. But nevertheless, as was shown, the parameters can be chosen in such a way that the specific character of the periodic variation of the coefficient yields an exponentially growing solution for $t\to\infty$ even when $\mu\neq 0$ is arbitrarily small. This phenomenon is known as **parametric resonance.** In particular, as is seen from the results of Sec. 3, this is the case for the equation $\ddot{y}+(1+\mu\cos 2t)\,y=0$.

This phenomenon can be demonstrated by a simple pendulum whose point of suspension is subject to forced vibrations along the vertical according to a given law $y=f(t)$. The equation describing the oscillations of such a pendulum (whose derivation is left to the reader) has the form $\ddot{\theta}+\left[\frac{g}{l}+\frac{f''(t)}{l}\right]\sin\theta=0$ where θ is the angular deviation from the vertical, g is the acceleration of gravity and l the length of the pendulum. Putting

$$\frac{g}{l}=\omega_0^2,\quad f(t)=M\cos\omega t,\quad \omega t=2t_1$$

and limiting ourselves to the linear approximation we arrive at the equation

$$\frac{d^2\theta}{dt_1^2}+\left[\left(\frac{2\omega_0}{\omega}\right)^2-\frac{4M}{l}\cos 2t_1\right]\theta=0 \tag{37}$$

(check it up!). We see that if $\frac{2\omega_0}{\omega}=1$, that is if the frequency of the forced vibrations of the point of suspension is twice the natural frequency of the pendulum, parametric resonance occurs, and the amplitude of oscillations grows. Another interesting phenomenon appears when the ratio $\frac{\omega_0}{\omega}$ is small: the upper equilibrium state of the pendulum (which of course can be realized if the pendulum is suspended by a weightless rod but not by a cord) may become stable while for the fixed point of suspension it is unstable. To obtain the linearized equation of small vibrations of the pendulum near its upper equilibrium position we must replace g by $-g$, i. e. take the term $\left(\frac{2\omega_0}{\omega}\right)^2$ in (37) with the minus sign. It is seen from Fig. 121 that the upper equilibrium state becomes stable if, for given M and l, the ratio $\frac{\omega_0}{\omega}$ is made sufficiently small, that is if the frequency of vibrations of the point of suspension is sufficiently high.

The notion of parametric resonance can also be elucidated by means of the following less rigorous but useful argument. Suppose that $\lambda > 0$ is finite (of the order of 1) while μ is small. It is readily seen that for $\mu = 0$ the expression

$$I = \frac{1}{2}(\dot{y}^2 + \lambda y^2)$$

is proportional to the total energy of the system and is independent of t. But if $\mu \neq 0$ we obtain the relation

$$\frac{dI}{dt} = -\mu [f_1(t) + \mu f_2(t) + \ldots] y\dot{y} \tag{38}$$

For small μ the solutions of equation (21) should be "close to harmonic oscillations", that is are expressible in the form $y = M \sin(\sqrt{\lambda}t + \varphi)$ where $M = M(t)$ and $\varphi = \varphi(t)$ are **slowly varying functions** in the sense that their rate of change vanishes for $\mu = 0$. From (38) it then follows that, to within the terms of higher order of smallness, we have

$$\frac{dI}{dt} = -\mu f_1(t) M \sin(\sqrt{\lambda}t + \varphi) M \sqrt{\lambda} \cos(\sqrt{\lambda}t + \varphi) =$$

$$= -\frac{\mu M^2 \sqrt{\lambda}}{2} \sin(2\sqrt{\lambda}t + 2\varphi) f(t) \tag{39}$$

The Fourier series of the function $f(t)$ only involving harmonics with frequences 2, 4, 6, ..., we see that if $\sqrt{\lambda}$ is not an integer the mean (average) value of the right member of (39) taken for a large time interval is close to zero, which indicates that the average rate of change of I is of higher order of smallness relative to μ. But if $\lambda = 1^2, 2^2, \ldots$, the mean value of the right member of (39) may be different from zero, that is the quantity I may vary at a rate of the order of μ. For instance, if $f(t) = \cos 2t$ and $\lambda = 1$ the average rate of change of I is equal to $-\frac{\mu M^2}{4} \sin \varphi$, and hence, depending on the value of the initial phase φ, the quantity I increases or decreases.

5. Hamiltonian Systems. Let us consider a canonical system of equations of form (VI.3.7) with a Hamiltonian function H which is a quadratic form with real coefficients:

$$H(t, \mathbf{y}, \mathbf{p}) = \frac{1}{2}\mathbf{y}^*\mathbf{P}(t)\mathbf{y} + \mathbf{p}^*\mathbf{Q}(t)\mathbf{y} + \frac{1}{2}\mathbf{p}^*\mathbf{R}(t)\mathbf{p},\ \mathbf{P}^* = \mathbf{P},\ \mathbf{R}^* = \mathbf{R} \tag{40}$$

Then the system is written as

$$\dot{\mathbf{y}} = \mathbf{Q}(t)\mathbf{y} + \mathbf{R}(t)\mathbf{p},\ \dot{\mathbf{p}} = -\mathbf{P}(t)\mathbf{y} - \mathbf{Q}^*(t)\mathbf{p} \tag{41}$$

and thus is linear. It sometimes is convenient to take this system in the form of the single vector equation

$$\frac{d}{dt}\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} = \mathbf{J}\begin{pmatrix}\mathbf{P}(t)\,\mathbf{Q}^*(t)\\ \mathbf{Q}(t)\,\mathbf{R}(t)\end{pmatrix}\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} \tag{42}$$

where

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} = \begin{pmatrix}y_1\\ \vdots\\ y_n\\ p_1\\ \vdots\\ p_n\end{pmatrix}$$ is a column vector and $\mathbf{J} = \begin{pmatrix}\mathbf{0} & \mathbf{I}\\ -\mathbf{I} & \mathbf{0}\end{pmatrix}$ is a square matrix

of order $2n$.

Let the reader prove that formulas (41) and (42) are equivalent by taking advantage of the following rule for multiplication of block matrices: if the blocks are such that their products entering into the resultant expression make sense, the operation can be performed by analogy with the ordinary rule of matrix multiplication, for instance,

$$\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}\begin{pmatrix}\mathbf{P} & \mathbf{Q}\\ \mathbf{R} & \mathbf{S}\end{pmatrix} = \begin{pmatrix}\mathbf{AP}+\mathbf{BR} & \mathbf{AQ}+\mathbf{BS}\\ \mathbf{CP}+\mathbf{DR} & \mathbf{CQ}+\mathbf{DS}\end{pmatrix}$$

where $\mathbf{AP}+\mathbf{BR}$, $\mathbf{AQ}+\mathbf{BS}$,' $\mathbf{CP}+\mathbf{DR}$ and $\mathbf{CQ}+\mathbf{DS}$ are the blocks forming the resultant matrix.

It can readily be shown that for any two solutions $\begin{pmatrix}\mathbf{y}_1\\ \mathbf{p}_1\end{pmatrix}$ and $\begin{pmatrix}\mathbf{y}_2\\ \mathbf{p}_2\end{pmatrix}$ of equation (42) we have the relation

$$\begin{pmatrix}\mathbf{y}_1\\ \mathbf{p}_1\end{pmatrix}^* \mathbf{J}\begin{pmatrix}\mathbf{y}_2\\ \mathbf{p}_2\end{pmatrix} \equiv \mathbf{y}_1^*\mathbf{p}_2 - \mathbf{p}_1^*\mathbf{y}_2 = \text{const}$$

To prove this property it suffices to differentiate the left-hand side with respect to t and then make use of equation (42) and the equality

$$\mathbf{J}^2 = -\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{pmatrix} = -\mathbf{J}^*\mathbf{J}$$

(let the reader carry out the calculations). Analogously, the solutions of the matrix equation

$$\dot{\mathbf{U}} = \mathbf{J}\begin{pmatrix}\mathbf{P}(t)\,\mathbf{Q}^*(t)\\ \mathbf{Q}(t)\,\mathbf{R}(t)\end{pmatrix}\mathbf{U} = \begin{pmatrix}\mathbf{Q}(t) & \mathbf{R}(t)\\ -\mathbf{P}(t) & -\mathbf{Q}^*(t)\end{pmatrix}\mathbf{U} \tag{43}$$

possess the same property. Here the unknown quantity $\mathbf{U}$ is a $(2n \times k)$-matrix function with an arbitrary number k of columns.

Now let us take a canonical system of form (42) with periodic coefficients of period T and prove the following **Lyapunov-Poincaré theorem:** *the characteristic equation of the monodromy matrix of periodic canonical system* (42) *is reciprocal.* (An algebraic equation of the form

$$f(\rho) \equiv a_0\rho^{2n} + a_1\rho^{2n-1} + \ldots + a_{2n-1}\rho + a_{2n} = 0 \qquad (44)$$

is said to be **reciprocal** if $a_0 = a_{2n}$, $a_1 = a_{2n-1}$, ..., or, in other words, if $\rho^{2n} f\left(\frac{1}{\rho}\right) \equiv f(\rho)$).

To prove the theorem let us take the $(2n \times 2n)$-matrix $\mathbf{U}(t)$ satisfying equation (43) and the initial condition $\mathbf{U}(0) = \mathbf{I}_{2n}$ where $\mathbf{I}_{2n}$ is the unit matrix of order $2n$. Then $\mathbf{U}(T)$ is the monodromy matrix whose eigenvalues are the characteristic multipliers. The property of the solutions of system (42) indicated above shows that $\mathbf{U}^*(t)\,\mathbf{J}\,\mathbf{U}(t) = \text{const}$, and therefore the initial condition implies

$$\mathbf{U}^*(t)\,\mathbf{J}\,\mathbf{U}(t) \equiv \mathbf{J}$$

Besides, the trace of the coefficient matrix on the right-hand side of (43) being equal to zero, we have, by virtue of (9), $\det \mathbf{U}(t) = \text{const}$, and hence the initial condition shows that

$$\det \mathbf{U}(t) \equiv 1$$

Consequently, if $f(\rho)$ is the characteristic polynomial of the matrix $\mathbf{U}(t)$ we can write (check up the calculations!)

$$f\left(\frac{1}{\rho}\right) = \det\left[\mathbf{U}(T) - \frac{1}{\rho}\mathbf{I}_{2n}\right] = \det\left[\mathbf{J}^{-1}(\mathbf{U}^*(T))^{-1}\mathbf{J} - \mathbf{J}^{-1}\frac{1}{\rho}\mathbf{I}_{2n}\mathbf{J}\right] =$$
$$= \det\left[(\mathbf{U}^*(T))^{-1} - \frac{1}{\rho}\mathbf{I}_{2n}\right] = \frac{1}{\rho^{2n}}\det\left[(\mathbf{U}^*(T))^{-1}(\rho\mathbf{I}_{2n} - \mathbf{U}^*(T))\right] =$$
$$= \frac{1}{\rho^{2n}}[\det \mathbf{U}^*(T)]^{-1}\det[\mathbf{U}^*(T) - \rho\mathbf{I}_{2n}] = \frac{1}{\rho^{2n}}f(\rho)$$

which is what we set out to prove.

If ρ is a root of a reciprocal equation the number $\frac{1}{\rho}$ is also its root, and therefore *all the solutions of a periodic Hamiltonian system cannot simultaneously tend to zero for* $t \to \infty$. Furthermore, *for all the solutions of such a system to be bounded it is necessary that the moduli of all the multipliers be equal to unity.*

Some interesting results related to parametric resonance (see Sec. 4) in Hamiltonian systems with more than one degree of freedom were obtained by M.G. Krein (see [92]). He considered

the system

$$\left.\begin{aligned}\dot{\mathbf{y}} &= \lambda \frac{\partial}{\partial \mathbf{p}}(H_0 + \mu H_1 + \mu^2 H_2 + \ldots)\\ \dot{\mathbf{p}} &= -\lambda \frac{\partial}{\partial \mathbf{y}}(H_0 + \mu H_1 + \mu^2 H_2 + \ldots)\end{aligned}\right\} \tag{45}$$

where all H_j's are quadratic forms in the variables $y_1, \ldots, y_n$, $p_1, \ldots, p_n$ with coefficients of period π in t, the form H_0 having constant coefficients and being positive definite. Let for $\lambda = 1$, $\mu = 0$ all the roots of the characteristic equation of the system be pairwise distinct and pure imaginary (denote them as $\pm i\omega_1$, $\pm i\omega_2, \ldots, \pm i\omega_n$), which means that all the solutions of the system are bounded for $-\infty < t < \infty$. Then, as in Sec. 3, there may appear regions of parametric resonance in the λ, μ-plane. Here we shall not dwell on the problem of determining the boundaries of these regions and only mention that the points at which such regions may adjoin the λ-axis must be of the form $\frac{2N}{\omega_j \pm \omega_k}$ where $N = 1, 2, 3, \ldots$. (Compare this assertion with the results of Sec. 3 and consider the problem of determining the frequences of parametric excitation for an autonomous Hamiltonian linear system.)

6. Nonhomogeneous Systems. For a nonhomogeneous system (2) we state the following properties of its solutions which are directly implied by formula (12):

1. *Let* $\|\mathbf{Y}(t, \tau)\| < \text{const}$ $(t_0 \leqslant \tau \leqslant t < \infty)$ *and* $\int_{t_0}^{\infty} \|\mathbf{f}(t)\|\,dt < \infty$.

Then all the solutions of system (2) *are bounded for* $t \to \infty$.

2. *Let* $\|\mathbf{Y}(t, \tau)\| \leqslant Ce^{-\alpha(t-\tau)}$ $(t_0 \leqslant \tau \leqslant t < \infty,\ \alpha > 0)$ *and* $\|\mathbf{f}(t)\| < \text{const}$. *Then all the solutions of system* (2) *are bounded for* $t \to \infty$.

3. *Let* $\|\mathbf{Y}(t, \tau)\| \leqslant Ce^{-\alpha(t-\tau)}$ $(t_0 \leqslant \tau \leqslant t < \infty,\ \alpha > 0)$ *and* $\|\mathbf{f}(t)\| \to 0$ *for* $t \to \infty$. *Then all the solutions of system* (2) *tend to zero as* $t \to \infty$.

To prove the last assertion (the proof of the first and the second is left to the reader) we note that under the hypotheses assumed we have

$$\left\|\int_{t_0}^{t} \mathbf{Y}(t, \tau)\mathbf{f}(\tau)\,d\tau\right\| \leqslant C\left[\int_{t_0}^{\frac{t+t_0}{2}} e^{-\alpha(t-\tau)}\|\mathbf{f}(\tau)\|\,d\tau + \int_{\frac{t+t_0}{2}}^{t} e^{-\alpha(t-\tau)}\|\mathbf{f}(\tau)\|\,d\tau\right] \leqslant$$

$$\leqslant \frac{C}{\alpha}\left[\max_{t_0 \leqslant \tau \leqslant \frac{t+t_0}{2}} \|\mathbf{f}(\tau)\| e^{-\frac{\alpha(t-t_0)}{2}} + \max_{\frac{t+t_0}{2} \leqslant \tau \leqslant t} \|\mathbf{f}(\tau)\|\right]$$

which implies the validity of the property (why?).

These propositions deal with properties common to all the solutions of system (2). There is also a variety of theorems related to the problem of existence of at least one solution of (2) possessing a certain property (such as boundedness, periodicity and the like). Here we shall limit ourselves to discussing the question of existence of a periodic solution. Let us begin with the system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{f}(t), \quad \mathbf{f}(t+T) \equiv \mathbf{f}(t) \tag{46}$$

whose coefficient matrix $\mathbf{A}$ is constant. Expanding the function $\mathbf{f}(t)$ and the sought-for solution $\mathbf{x}(t)$ into Fourier's series

$$\mathbf{f}(t) = \sum_k e^{ik\omega t}\mathbf{f}_k, \quad \mathbf{x}(t) = \sum_k e^{ik\omega t}\mathbf{x}_k \left(\omega = \frac{2\pi}{T}\right)$$

substituting these series into (46) and equating the coefficients in the same exponential expressions we obtain the relations

$$(\mathbf{A} - ik\omega\mathbf{I})\,\mathbf{x}_k = -\mathbf{f}_k \qquad (k = 0, \pm 1, \pm 2, \ldots)$$

Thus, system (46) possesses exactly one periodic solution if none of the numbers $ik\omega$ is an eigenvalue of the matrix $\mathbf{A}$. If this condition is violated for some $k = k_0$ the vector $\mathbf{f}_{k_0}$ must satisfy the corresponding orthogonality conditions because, if otherwise, there appear resonant terms and the periodic solution does not exist. If the orthogonality conditions are fulfilled the solution is determined to within a term of the form $e^{ik_0\omega t}\mathbf{x}_{k_0}$ where $\mathbf{x}_{k_0}$ is any eigenvector of the matrix $\mathbf{A}$ belonging to the eigenvalue $ik_0\omega$. Let the reader specify the orthogonality conditions mentioned here.

Consider now a nonhomogeneous periodic system of the form

$$\dot{\mathbf{x}} = \mathbf{A}(t)\,\mathbf{x} + \mathbf{f}(t), \quad \mathbf{A}(t+T) \equiv \mathbf{A}(t), \quad \mathbf{f}(t+T) \equiv \mathbf{f}(t) \tag{47}$$

This system remains invariant under the transformation $t \to t+T$. Consequently, if $\mathbf{x}(t)$ is a solution of the system, the function $\mathbf{x}(t+T)$ is also its solution. It follows that if a solution of system (47) satisfies the condition

$$\mathbf{x}(0) = \mathbf{x}(T) \tag{48}$$

then $\mathbf{x}(t) \equiv \mathbf{x}(t+T)$. Indeed, both members of equality (48) satisfy system (47) and coincide for $t = 0$, which means that they coincide identically (why?).

By virtue of formula (12), the periodicity condition (48) can be rewritten in the form

$$\mathbf{x}(0) = \int_0^T \mathbf{Y}(T, \tau)\,\mathbf{f}(\tau)\,d\tau + \mathbf{Y}(T, 0)\,\mathbf{x}(0) \tag{49}$$

Thus, if 1 is not an eigenvalue of the monodromy matrix $\mathbf{Y}(T, 0)$, equation (49) has a uniquely determined solution $\mathbf{x}(0)$, that is system (47) possesses exactly one periodic solution with period T for any T-periodic function $\mathbf{f}(t)$. But if 1 is an eigenvalue of the monodromy matrix, the integral in formula (49) must satisfy the corresponding orthogonality condition which, ultimately, is a condition on $\mathbf{f}(t)$. If this condition holds the initial value $\mathbf{x}(0)$ is determined to within an arbitrary eigenvector of the matrix $\mathbf{Y}(T, 0)$ associated with the eigenvalue 1. If the monodromy matrix possesses the eigenvalue 1, the homogeneous system corresponding to system (47) has nonzero periodic solutions of period T. It can be shown (let the reader elaborate the proof) that if in the latter case all the solutions of the homogeneous system are bounded for $t \longrightarrow \infty$ while the above orthogonality conditions are violated, the solution $\mathbf{x}(t)$ of system (47) becomes unbounded for $t \longrightarrow \infty$ and grows at the rate of a power function, that is resonance occurs.

For systems (46) and (47) the following criterion (whose proof we leave to the reader) is sometimes used: *for at least one periodic solution with period T to exist it is necessary and sufficient that the equality* $\int_0^T (\mathbf{z}(t), \mathbf{f}(t))\, dt = 0$ *be fulfilled for every T-periodic solution $\mathbf{z}(t)$ of the corresponding adjoint system* (6).

Let us consider a class of nonhomogeneous systems known as systems with **pulse excitation**. The solution $\mathbf{x}(t)$ of such a system gains instantaneous increments ("jumps") $\Delta_j \mathbf{x}$ at given time moments t_j, $j = 1, 2, \ldots$ $(t_0 < t_1 < t_2 < \ldots \longrightarrow \infty)$ and satisfies equation (2) in the intervals between the pulses. In the simplest case the quantities $\Delta_j \mathbf{x}$ are given but in the general case they may depend on the sought-for solution. Using the notion of a delta function (e.g. see [107], Sec. XIV.25) we can write a system of this type with given $\Delta_j \mathbf{x}$'s in the form

$$\dot{\mathbf{x}} = \mathbf{A}(t)\,\mathbf{x} + \mathbf{f}(t) + \sum_j \Delta_j \mathbf{x}\, \delta(t - t_j)$$

Therefore, the solution can be obtained by means of formula (12) into which the modified nonhomogeneity term should be substituted. On the basis of the properties of integrals involving delta functions (e.g. see [107], formula (XIV.102)) we then obtain

$$\mathbf{x}(t) = \int_{t_0}^{t} \mathbf{Y}(t, \tau)\, \mathbf{f}(\tau)\, d\tau + \mathbf{Y}(t, t_0)\, \mathbf{x}_0 + \sum_{t_j < t} \mathbf{Y}(t, t_j)\, \Delta_j \mathbf{x}$$

This formula which can also be derived with the help of the superposition principle (how?) can be used for investigating the properties of the system.

7. Almost Periodic Functions. The so-called **almost periodic functions** are an important generalization of the class of periodic functions. An almost periodic function is a superposition of periodic functions whose frequencies are, in the general case, incommensurable. If $f_1(t)$ and $f_2(t)$ are two periodic functions whose periods T_1 and T_2 are commensurable, that is are in the ratio $m_1 : m_2$ where m_1 and m_2 are integers, then $T_1 = m_1 T$, $T_2 = m_2 T$, and the sum of the functions and also their product, etc. are periodic functions with period $m_1 m_2 T$. But if the periods T_1 and T_2 are incommensurable the sum $f_1(t) + f_2(t)$ is no longer periodic. For instance, such is the function $\sin t + \sin \sqrt{2} t$ which is not periodic and belongs to a wider class of functions referred to as almost periodic.

An almost periodic function can be defined as one which can be uniformly approximated on the whole t-axis, with an arbitrary accuracy, by finite sums of the form

$$f(t) = \sum_j c_j e^{i\omega_j t} \tag{50}$$

where ω_j are real numbers and c_j are some complex coefficients. In particular, the sum of an arbitrary series of form (50) whose coefficients satisfy the condition

$$\sum_j |c_j| < \infty \tag{51}$$

is always an almost periodic function. According to Weierstrass' test (e.g. see [107], Sec. XVII.8), condition (51) guarantees uniform convergence of series (50) on the entire t-axis. (In this connection we note that other types of deviation of functions can also be applied; this leads to conditions weaker than (51), which makes it possible to study discontinuous almost periodic functions as well.) Various classes of such functions and their properties were considered in [13] by H. Bohr (1887-1951), a Danish mathematician, who inaugurated the theory of almost periodic functions in 1923. These questions are also treated in [85].

Here we give without proof some properties of continuous almost periodic functions: a sum, a product and a quotient of almost periodic functions (in the case of a quotient it is naturally supposed that the denominator cannot be arbitrarily close to zero for $-\infty < t < \infty$) are again almost periodic functions. Furthermore, the limit of an arbitrary sequence of almost periodic functions which is uniformly convergent on the whole t-axis is also a function of that class. Finally, for a continuous function $f(t)$ to be almost periodic it is necessary and sufficient that every sequence of the form $\{f(t + \tau_n)\}$ contain a uniformly convergent subsequence. From the last property it follows, in particular, that if $f(t) \not\equiv \text{const}$ then this function cannot tend to a constant for $t \to \infty$. It should

be stressed that every periodic function belongs to the class of almost periodic functions but the converse is not true, that is an almost periodic function may not be periodic.

Every almost periodic function is bounded on the whole t-axis and possesses the mean value

$$\overline{f(t)} = \lim_{T \to \infty} \frac{1}{2T} \int_{\alpha - T}^{\alpha + T} f(t)\, dt$$

independent of α. We have

$$\overline{e^{i\omega t}} = \begin{cases} 0 & \text{for } \omega \neq 0 \\ 1 & \text{for } \omega = 0 \end{cases}$$

(check it up!), and therefore from representation (50) it follows that

$$\overline{f(t) e^{-i\omega t}} = \begin{cases} c_j \text{ if } \omega \text{ coincides with one of the numbers } \omega_j, \\ 0 \text{ for all the other } \omega\text{'s} \end{cases} \tag{52}$$

Consequently, the set of the frequencies ω_j (which is referred to as the **spectrum** of the almost periodic function $f(t)$) and the corresponding **amplitudes** c_j can be determined by computing the mean values $\overline{f(t) e^{-i\omega t}}$. Relation (50) also implies the formula

$$\overline{|f(t)|^2} = \overline{f(t)\,[f(t)]^*} = \overline{\sum_{j,k} c_j c_k^* e^{i(\omega_j - \omega_k)t}} = \sum_j |c_j|^2$$

which turns into the Parseval relation (e.g. see [107], Sec. XVII.27) when $f(t)$ is a periodic function.

Real almost periodic functions possess the following characteristic property (which readily follows from (52): if the spectrum of such a function contains a frequency ω_j, the frequency $\omega_{-j} = -\omega_j$ also enters into the spectrum, the corresponding amplitudes satisfying the relation $c_{-j} = c_j^*$.

Now let us consider a system of differential equations with constant coefficients and almost periodic nonhomogeneity term:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{f}(t), \quad \mathbf{f}(t) = \sum_j e^{i\omega_j t} \mathbf{c}_j \tag{53}$$

If we seek a solution of the system as an almost periodic function with the same spectrum, that is in the form $\mathbf{x}(t) = \sum_j e^{i\omega_j t} \mathbf{x}_j$, the substitution of this expression into (53) yields the relations

$$(\mathbf{A} - i\omega_j \mathbf{I})\mathbf{x}_j = -\mathbf{f}_j$$

If none of the numbers $i\omega_j$ is an eigenvalue of the matrix $\mathbf{A}$ the sought-for solution must be of the form

$$\mathbf{x}(t) = -\sum_j e^{i\omega_j t} (\mathbf{A} - i\omega_j \mathbf{I})^{-1} \mathbf{f}, \tag{54}$$

If the matrix **A** has no pure imaginary eigenvalues, formula (54) in fact represents a solution of (53); moreover, we see that in this case system (53) possesses exactly one almost periodic solution given by formula (54). However, if there are eigenvalues of the form $i\tilde{\omega}$ (where $\tilde{\omega}$ does not necessarily enter into the spectrum), among the numbers ω_j there may be frequencies which are arbitrarily close to $\tilde{\omega}$, and then the norm $\|(\mathbf{A}-i\omega_j\mathbf{I})^{-1}\|$ may become arbitrarily large (why?). In this case series (54) may be divergent unless the increase of the terms of the series is compensated for by an adequate decrease of the coefficients $\mathbf{f}_j$.

In applications we most often deal with almost periodic functions whose spectrum results from a superposition of vibrations with a finite number of incommensurable frequencies, that is contains frequencies of the form $\omega = m_1\omega_1 + m_2\omega_2 + \ldots + m_p\omega_p$ where the numbers $\omega_1, \omega_2, \ldots, \omega_p$ are given and $m_1, m_2, \ldots, m_p$ are arbitrary integers. Such functions are sometimes referred to as **quasi-periodic** but it should be noted that this term is not commonly recognized and is sometimes used as a synonym for an almost periodic function of the general type. P. Bohl was the first to investigate such functions in 1893. In particular, he showed that every function of this type can be represented in the form $f(\omega_1 t, \omega_2 t, \ldots, \omega_p t)$ where $f(s_1, s_2, \ldots, s_p)$ is a function periodic in each of its arguments with period 2π (the converse is also true: it is obvious that expanding a function $f(s_1, s_2, \ldots, s_p)$ into Fourier's p-fold series and replacing s_j by $\omega_j t$ $(j = 1, 2, \ldots, p)$ we obtain a quasi-periodic function).

8. **Asymptotic Expansions of Solutions for $t \to \infty$.** There are many interesting results related to the asymptotic behaviour of solutions of system (4) for $t \to \infty$. Here we present some of them. For more detail we refer the reader to [8], [18], [20], [57] and [137].

To elucidate the character of the problem we begin with the simplest case of one scalar equation $(n = 1)$

$$\dot{y} = \left(a_0 + \frac{a_1}{t} + \frac{a_2}{t^2} + \ldots\right) t^k y$$

where $k \geqslant 0$ is an integer. Separating the variables, integrating and taking the antilogarithms we obtain

$$y = C \exp\left(\frac{a_0}{k+1} t^{k+1} + \frac{a_1}{k} t^k + \ldots + a_k t\right) t^{a_{k+1}} \exp\left(-\frac{a_{k+2}}{t} - \frac{a_{k+3}}{2t^2} \ldots\right)$$

Substituting the series expansion of the second exponential expression taken for large t into this formula and combining similar terms we arrive at the representation

$$y = C \exp\left(\frac{a_0}{k+1} t^{k+1} + \ldots + a_k t\right) t^{a_{k+1}} \left(1 + \frac{b_1}{t} + \frac{b_2}{t^2} + \ldots\right)$$

where $b_1, b_2, \ldots$ are some constant coefficients.

For $k=-1$ we obtain an analogous formula without the first exponential expression. Finally, for $k<-1$ we obtain the formula

$$y=C\left(1+\frac{b_1}{t-(k+2)+1}+\frac{b_2}{t-(k+2)+2}+\ldots\right)$$

(check it up!).

It turns out that in the general case of an arbitrary number n of equations the expansions of the solutions are analogous to the above but, generally speaking, for $k\geqslant 0$ the series thus obtained are only asymptotically convergent (see Secs. II.4.1 and 2). Let us consider the system

$$\dot{\mathbf{y}}=\mathbf{A}(t)\,t^k\mathbf{y} \tag{55}$$

where $k=0, 1, 2, \ldots$ and $\mathbf{A}(t)=\mathbf{A}_0+\frac{\mathbf{A}_1}{t}+\frac{\mathbf{A}_2}{t^2}+\ldots$ (the latter equality may have an asymptotic sense). It can be proved that to each eigenvalue λ_j of the matrix $\mathbf{A}_0$ there corresponds a solution of system (55) expressible in the form

$$\mathbf{y}=\exp\left(\frac{\lambda_j}{k+1}t^{k+1}+\alpha_1 t^k+\ldots+\alpha_k t\right)t^{\beta}\left(\mathbf{b}_0+\frac{\mathbf{b}_1}{t}+\frac{\mathbf{b}_2}{t^2}+\ldots\right) \tag{56}$$

where, in the general case, the series in the parentheses converges to its sum asymptotically. The constant quantities $\alpha_1, \ldots, \alpha_k$, β, $\mathbf{b}_0$, $\mathbf{b}_1$, $\mathbf{b}_2$, ... can be found by means of the method of undetermined coefficients if we substitute (56) into (55). After the cancellation this yields the relation

$$\begin{aligned}&(\lambda_j t^k+\alpha_1 k t^{k-1}+\ldots+\alpha_k)\left(\mathbf{b}_0+\frac{\mathbf{b}_1}{t}+\frac{\mathbf{b}_2}{t^2}+\ldots\right)+\\&+\beta\left(\frac{\mathbf{b}_0}{t}+\frac{\mathbf{b}_1}{t^2}+\frac{\mathbf{b}_2}{t^3}+\ldots\right)+\left(-\frac{\mathbf{b}_1}{t^2}-2\frac{\mathbf{b}_2}{t^3}-3\frac{\mathbf{b}_3}{t^4}-\ldots\right)=\\&=t^k\left(\mathbf{A}_0+\frac{\mathbf{A}_1}{t}+\frac{\mathbf{A}_2}{t^2}+\ldots\right)\left(\mathbf{b}_0+\frac{\mathbf{b}_1}{t}+\frac{\mathbf{b}_2}{t^2}+\ldots\right)\end{aligned} \tag{57}$$

whence, opening the brackets and equating the coefficients in the same power of t we find the desired quantities.

For definiteness, let us suppose that $k=1$ and that the eigenvalue λ_j is simple. Then the above procedure results in the equalities

$$\left.\begin{aligned}&\mathbf{A}_0\mathbf{b}_0-\lambda_j\mathbf{b}_0=0,\quad \mathbf{A}_0\mathbf{b}_1-\lambda_j\mathbf{b}_1=\alpha_1\mathbf{b}_0-\mathbf{A}_1\mathbf{b}_0\\&\mathbf{A}_0\mathbf{b}_2-\lambda_j\mathbf{b}_2=\alpha_1\mathbf{b}_1-\mathbf{A}_1\mathbf{b}_1+\beta\mathbf{b}_0-\mathbf{A}_2\mathbf{b}_0\\&\mathbf{A}_0\mathbf{b}_3-\lambda_j\mathbf{b}_3=\alpha_1\mathbf{b}_2-\mathbf{A}_1\mathbf{b}_2+\beta\mathbf{b}_1-\mathbf{A}_2\mathbf{b}_1-\mathbf{b}_1-\mathbf{A}_3\mathbf{b}_0,\ \ldots\end{aligned}\right\} \tag{58}$$

From the first equality we conclude that $\mathbf{b}_0$ is an eigenvector of the matrix $\mathbf{A}_0$ belonging to the eigenvalue λ_j. Fixing $\mathbf{b}_0$ we then can determine α_1 by applying the condition of the orthogonality of the right member of the second equality to a vector $\mathbf{d}\neq 0$ for which $\mathbf{A}_0^*\mathbf{d}=\lambda_j^*\mathbf{d}$ (in this connection remember the solvability con-

ditions for an equation of the form $\mathbf{Ax}=\mathbf{b}$). This orthogonality condition is

$$\alpha_1(\mathbf{b}_0,\ \mathbf{d})-(\mathbf{A}_1\mathbf{b}_0,\ \mathbf{d})=0 \tag{59}$$

(it can readily be shown that if λ_1 is a simple eigenvalue we have $(\mathbf{b}_0,\ \mathbf{d})\neq 0$). If condition (59) is fulfilled the solution of the second equality (58) involves one degree of freedom, that is we obtain $\mathbf{b}_1=\mathbf{b}_1'+C_1\mathbf{b}_0$ where C_1 is an arbitrary parameter. The orthogonality condition for the third equality (58) can be written, by virtue of (59), in the form

$$\alpha_1(\mathbf{b}_1',\ \mathbf{d})-(\mathbf{A}_1\mathbf{b}_1',\ \mathbf{d})+\beta(\mathbf{b}_0,\ \mathbf{d})-(\mathbf{A}_2\mathbf{b}_0,\ \mathbf{d})=0 \tag{60}$$

and consequently it enables us to determine β, after which we obtain $\mathbf{b}_2=\mathbf{b}_2'+C_2\mathbf{b}_0$. Next, taking the orthogonality condition for the fourth equality (58) we use conditions (59) and (60) and thus determine C_1 in a unique manner (check it up!). Then we obtain $\mathbf{b}_3=\mathbf{b}_3'+C_3\mathbf{b}_0$, determine C_2 from the subsequent equality and so on.

In this way we can obtain asymptotic expansions for all solutions of system (55) if all the eigenvalues of the matrix $\mathbf{A}_0$ are simple (this also applies to a more general case when the matrix $\mathbf{A}_0$ only involves Jordan submatrices of the first order; see Sec. IV.3.3). The construction of asymptotic expansions in the case of Jordan blocks of higher order and the proof that these expansions not only formally satisfy system (55) but in fact yield asymptotic expansions of solutions is connected with more sophisticated considerations for which we refer the reader to the books indicated above.

As an example, let us consider Bessel's equation (e.g. see [107], Sec. XV.26):

$$x^2y''+xy'+(x^2-p^2)y=0 \qquad (0<x<\infty) \tag{61}$$

Putting $y=y_1$ and $y'=y_2$ we pass from equation (61) to the system

$$\left.\begin{aligned} y_1'&=y_2\\ y_2'&=-\frac{1}{x^2}[xy_2+(x^2-p^2)y_1]=\left(-1+\frac{p^2}{x^2}\right)y_1-\frac{1}{x}y_2 \end{aligned}\right\} \tag{62}$$

We have obtained a system of form (55) with $k=0$ whose matrix is

$$\mathbf{A}_0=\begin{pmatrix} 0 & 1\\ -1 & 0\end{pmatrix}$$

its eigenvalues being $\lambda_{1,2}=\pm i$. Choosing λ_1 we conclude, on the basis of (56), that the asymptotic expansion should be looked for in the form

$$y=e^{ix}x^{\beta}\left(b_0+\frac{b_1}{x}+\frac{b_2}{x^2}+\ldots\right) \qquad (x\to\infty)$$

The direct substitution of this expression into (61) (which is equivalent to the substitution of the corresponding expressions of y_1

and y_2 into (62)) yields, after the coefficients in like powers of x have been equated, the equalities (check it up!)

$$2b_0\beta i+b_0 i=0,\ b_0\beta(\beta-1)+2b_1(\beta-1)i+b_0\beta+b_1 i-p^2b_0=0$$
$$b_1\beta(\beta-3)+2b_2(\beta-2)i+b_1(\beta-1)+b_2 i-p^2b_1=0,\ \ldots$$

from which we find, in succession,

$$\beta=-\frac{1}{2},\ b_1=\frac{i}{2}\left(p^2-\frac{1}{4}\right)b_0,\ b_2=-\frac{1}{8}\left(p^2-\frac{1}{4}\right)^2 b_0,\ \ldots$$

Thus we arrive at the asymptotic expansion

$$y\sim b_0 e^{ix}\frac{1}{\sqrt{x}}\left\{1+\frac{i}{2}\left(p^2-\frac{1}{4}\right)\frac{1}{x}-\frac{1}{8}\left(p^2-\frac{1}{4}\right)^2\frac{1}{x^2}+\ldots\right\}$$

The comparison with formula (II.4.26) shows that if we put $b_0=\sqrt{\frac{2}{\pi}}\exp\left[-i\left(\frac{p\pi}{2}+\frac{\pi}{4}\right)\right]$ the asymptotic expansion will represent the function $H_p^{(1)}(x)$. Separating the real and imaginary parts we can obtain the asymptotic expansion for the functions $J_p(x)$ and $Y_p(x)$.

If $k=-1$, then, instead of (56), we use the formula

$$\mathbf{y}=t^{\beta}\left(\mathbf{b}_0+\frac{\mathbf{b}_1}{t}+\frac{\mathbf{b}_2}{t^2}+\ldots\right) \tag{63}$$

for the sought-for solution. For the coefficients of the expansion we obtain in this case, instead of (57), the relation

$$\left[(\mathbf{A}_0-\beta\mathbf{I})+\frac{\mathbf{A}_1}{t}+\frac{\mathbf{A}_2}{t^2}+\ldots\right]\left(\frac{\mathbf{b}_0}{t}+\frac{\mathbf{b}_1}{t^2}+\frac{\mathbf{b}_2}{t^3}+\ldots\right)=$$
$$=-\frac{\mathbf{b}_1}{t^2}-2\frac{\mathbf{b}_2}{t^3}-3\frac{\mathbf{b}_3}{t^4}-\ldots$$

(check it up!). Opening the brackets we then arrive at the equalities

$$(\mathbf{A}_0-\beta\mathbf{I})\mathbf{b}_0=\mathbf{0},\ [\mathbf{A}_0-(\beta-1)\mathbf{I}]\ \mathbf{b}_1=-\mathbf{A}_1\mathbf{b}_0$$
$$[\mathbf{A}_0-(\beta-2)\mathbf{I}]\mathbf{b}_2=-\mathbf{A}_2\mathbf{b}_0-\mathbf{A}_1\mathbf{b}_1-\ldots$$

From the first of these equalities we see that β is an eigenvalue of the matrix $\mathbf{A}_0$ while $\mathbf{b}_0$ is an eigenvector belonging to this eigenvalue. After β and $\mathbf{b}_0$ have been chosen, we obtain from the second equality, provided that $\beta-1$ is not an eigenvalue of the matrix $\mathbf{A}_0$, the vector $\mathbf{b}_1$. Proceeding in this manner we then take the third equality from which, if $\beta-2$ is not an eigenvalue, we determine $\mathbf{b}_2$, etc. We see that there appear singular cases (which shall not be treated here) when $\beta-m$ is an eigenvalue of the matrix $\mathbf{A}_0$ for an integer $m>0$. In the simpler case when the numbers $\beta-m$ are not eigenvalues and series (55) for $\mathbf{A}(t)$ is convergent in the ordinary (not asymptotic) sense it can be proved (e.g. see the books mentio-

ned above) that the resultant series (63) converges in the ordinary sense for sufficiently large values of t, and thus we obtain the ordinary series expansion for the sought-for solution.

If $k \leqslant -2$ the solution should be looked for in the form

$$\mathbf{y} = \mathbf{b}_0 + \mathbf{b}_1 t^{k+1} + \mathbf{b}_2 t^k + \mathbf{b}_3 t^{k-1} + \ldots$$

where $\mathbf{b}_0$ is an arbitrary vector. (Let the reader derive the relations for determining the coefficients $\mathbf{b}_1$, $\mathbf{b}_2$, ...) In this case the series obtained is also convergent to the solution in the ordinary sense.

9. More on Asymptotic Behaviour of Solutions. Unfortunately, the derivation of an asymptotic expansion which yields valuable information on the behaviour of the solutions is by far not always available (for instance, it is connected with essential difficulties when the coefficients of system (4) are periodic, and the like). Therefore, it sometimes is expedient to resort to some propositions of qualitative character. Below are two examples of this kind for which the following lemma is needed:

Lemma. *Let a function $u(t)$ satisfy the inequality*

$$u(t) \leqslant \int_{t_0}^{t} \rho(\tau)\, u(\tau)\, d\tau + C \quad (t_0 \leqslant t \leqslant t_1) \tag{64}$$

where $\rho(t) \geqslant 0$. Then $u(t) \leqslant C \exp \int_{t_0}^{t} \rho(\tau)\, d\tau$ for $t_0 \leqslant t \leqslant t_1$.

Indeed, since $\rho \geqslant 0$ we have

$$\rho(t)\, u(t) \leqslant \rho(t) \left[\int_{t_0}^{t} \rho(\tau)\, u(\tau)\, d\tau + C \right]$$

and hence, by virtue of (64), we obtain

$$u(t) \leqslant \int_{t_0}^{t} \rho(\tau) \left[\int_{t_0}^{\tau} \rho(\tau_1)\, u(\tau_1)\, d\tau_1 + C \right] d\tau + C$$

which means that inequality (64) can be iterated. Performing successive iterations of this type and passing to the limit we obtain (see Sec. VII.2.9) the limiting right member (denote it by $v(t)$) satisfying the integral equation

$$v(t) = \int_{t_0}^{t} \rho(\tau)\, v(\tau)\, d\tau + C$$

(check it up!). The differentiation readily shows that the solution of this equation is equal to $C \exp \int_{t_0}^{t} \rho(\tau)\, d\tau$, which implies the assertion of the lemma.

The theorems below make it possible to judge upon the asymptotic behaviour of a given system of the form

$$\dot{\mathbf{z}} = [\mathbf{A}(t) + \mathbf{B}(t)]\,\mathbf{z} \tag{65}$$

when the asymptotic behaviour of the solutions of the corresponding system (4) (with matrix $\mathbf{A}(t)$) is known (for instance, in the case of a constant or periodic coefficient matrix) under the assumption that the matrix function $\mathbf{B}(t)$ is in a certain sense small for $t \longrightarrow \infty$. A number of results of this kind can be found in [8].

1. *Let*

$$\| \mathbf{Y}(t, \tau) \| < \text{const} \quad (t_0 \leqslant \tau \leqslant t < \infty), \quad \int_{t_0}^{\infty} \| \mathbf{B}(t) \|\, dt < \infty \tag{66}$$

(for simplicity, the norm of a matrix can be understood here, unlike Sec. IV.4.2, as the sum of the moduli of its elements). *Then all the solutions of system* (65) *are bounded for* $t \longrightarrow \infty$.

To prove the theorem, let us substitute into equation (65) its solution $\mathbf{z}(t)$ satisfying an arbitrary initial condition $\mathbf{z}(t_0) = \mathbf{z}_0$, open the square brackets on the right-hand side of (65) and, considering the expression $\mathbf{B}(t)\,\mathbf{z}$ as a nonhomogeneity term, apply formula (12). This yields

$$\mathbf{z}(t) = \int_{t_0}^{t} \mathbf{Y}(t, \tau)\, \mathbf{B}(\tau)\, \mathbf{z}(\tau)\, d\tau + \mathbf{Y}(t, t_0)\, \mathbf{z}_0 \tag{67}$$

However, it can readily be verified directly that the norm we consider here possesses the properties

$$\| \mathbf{A} + \mathbf{B} \| \leqslant \| \mathbf{A} \| + \| \mathbf{B} \|, \quad \| \mathbf{AB} \| \leqslant \| \mathbf{A} \| \cdot \| \mathbf{B} \|$$
$$\| \mathbf{Ax} \| \leqslant \| \mathbf{A} \| \cdot \| \mathbf{x} \|, \quad \left\| \int \mathbf{x}(t)\, dt \right\| \leqslant \int \| \mathbf{x}(t) \|\, dt$$

where the norm of a number vector is understood as the sum of the absolute values of its components. Consequently, relation (67) and the hypotheses of the theorem imply that

$$\| \mathbf{z}(t) \| \leqslant C_1 \int_{t_0}^{t} \| \mathbf{B}(\tau) \| \cdot \| \mathbf{z}(\tau) \|\, d\tau + C_2 \| \mathbf{z}_0 \| \qquad (t_0 \leqslant t < \infty) \tag{68}$$

where C_1 and C_2 are some constants. Now, applying the above lemma, we derive the inequality

$$\| \mathbf{z}(t) \| \leqslant C_2 \| \mathbf{z}_0 \| \exp \left[C_1 \int_{t_0}^{t} \| \mathbf{B}(\tau) \|\, d\tau \right] \qquad (t_0 \leqslant t < \infty)$$

which is what we wished to prove.

In particular, it should be noted that the first condition (66) is always fulfilled for systems with constant or periodic coefficients whose all solutions are bounded. Besides, this condition does not depend on the choice of t_0 (prove it!).

2. *Let*

$$\| \mathbf{Y}(t, \tau) \| \leqslant Ce^{-\alpha(t-\tau)} \qquad (t_0 \leqslant \tau \leqslant t < \infty)$$

and

$$\| \mathbf{B}(t) \| \leqslant \varepsilon < \frac{\alpha}{C} \qquad (t_0 \leqslant t < \infty) \tag{69}$$

where C, α *and* ε *are positive constants. Then all the solutions of system* (65) *tend to zero as* $t \longrightarrow \infty$.

Actually, from (67), by analogy with (68), we readily derive

$$\| \mathbf{z}(t) \| \leqslant C\varepsilon \int_{t_0}^{t} e^{-\alpha(t-\tau)} \| \mathbf{z}(\tau) \| d\tau + Ce^{-\alpha(t-t_0)} \| \mathbf{z}_0 \|$$

whence

$$\| \mathbf{z}(t) \| e^{\alpha t} \leqslant C\varepsilon \int_{t_0}^{t} \| \mathbf{z}(\tau) \| e^{\alpha\tau} d\tau + Ce^{\alpha t_0} \| \mathbf{z}_0 \|$$

Applying the lemma proved at the beginning of this section we conclude that

$$\| \mathbf{z}(t) \| e^{\alpha t} \leqslant Ce^{\alpha t_0} \| \mathbf{z}_0 \| e^{C\varepsilon(t-t_0)} \qquad (t_0 \leqslant t < \infty)$$

whence the assertion of the theorem follows.

The first condition (69) is, in particular, always fulfilled for systems with constant or periodic coefficients whose all solutions tend to zero for $t \longrightarrow \infty$. Besides, it suffices to require that condition (69) should only hold for sufficiently large values of t, and therefore we can simply impose the condition that $\mathbf{B}(t) \underset{t \to \infty}{\longrightarrow} 0$.

At the same time it should be stressed that, generally speaking, the boundedness of the solutions of system (4) and the condition $\mathbf{B}(t) \underset{t \to \infty}{\longrightarrow} 0$ do not imply the boundedness of the solutions of system (65). This is confirmed by the following simple example: the solutions of the scalar equation $\dot{y} = 0y$ are bounded while those of the equation $\dot{z} = \left(0 + \frac{1}{t}\right) z$ are not.

Here we also give another useful estimation for the solutions of system (4):

$$|\mathbf{y}(t_0)| \exp \int_{t_0}^{t} \lambda_{\min}(\tau)\, d\tau \leqslant |\mathbf{y}(t)| \leqslant |\mathbf{y}(t_0)| \exp \int_{t_0}^{t} \lambda_{\max}(\tau)\, d\tau$$

where $\lambda_{\min}(t)$ and $\lambda_{\max}(t)$ are, respectively, the least and the greatest

eigenvalues of the Hermitian matrix $\frac{1}{2}[\mathbf{A}(t)+\mathbf{A}^*(t)]$. To prove the estimation we take advantage of the relation

$$\frac{d}{dt}(|\mathbf{y}|^2)=\frac{d}{dt}(\mathbf{y}^*\mathbf{y})=\dot{\mathbf{y}}^*\mathbf{y}+\mathbf{y}^*\dot{\mathbf{y}}=\mathbf{y}^*\mathbf{A}^*\mathbf{y}+\mathbf{y}^*\mathbf{A}\mathbf{y}=\mathbf{y}^*(\mathbf{A}+\mathbf{A}^*)\mathbf{y}$$

whence, by the extremal property of the eigenvalues, it follows that

$$\lambda_{\min}(t)\leqslant\frac{\frac{d}{dt}(|y|^2)}{2|y|^2}\leqslant\lambda_{\max}(t), \text{ etc.}$$

10. Oscillation of Solutions of Second-Order Equations. Linear differential equations of the form

$$y''+a(x)y'+b(x)y=0 \tag{70}$$

(here the independent variable is denoted by x) are widely encountered in various divisions of theoretical and applied mathematics. The properties of these equations are most thoroughly investigated.

We shall begin with a simplification that can be achieved in equation (70) by means of a linear substitution

$$y=\alpha(x)z \tag{71}$$

where the function $\alpha(x)\neq 0$ is chosen in a certain manner. Substituting (71) into (70) we obtain the equation for z:

$$z''+\frac{1}{\alpha}[2\alpha'+a(x)\alpha]z'+\frac{1}{\alpha}[\alpha''+a(x)\alpha'+b(x)\alpha]z=0 \tag{72}$$

Now, in particular, we see that it is possible to eliminate the term with z'; for this purpose it suffices to choose as $\alpha(x)$ a nonzero solution of the equation $2\alpha'+a(x)\alpha=0$, whence, to within an inessential nonzero constant factor, $\alpha=\exp\left(-\frac{1}{2}\int_{x_0}^{x}a(\xi)d\xi\right)$. The substitution of the function $\alpha(x)$ thus determined into (72) leads to the equation

$$z''+\left[b(x)-\frac{1}{2}a'(x)-\frac{1}{4}a^2(x)\right]z=0 \tag{73}$$

(let the reader carry out the calculations!).

There is another method of elimination of the term involving the first derivative based on a change of the independent variable

$$\xi=\varphi(x) \tag{74}$$

where the function $\varphi(x)$ should be appropriately chosen. (Note that a change $\eta=\psi(y)$ of the unknown function which is analogous to (74) would lead to a nonlinear equation for η.) Substituting (74)

into (70) we obtain for $y(\xi)$ the equation

$$[\varphi'(x)]^2 \frac{d^2y}{d\xi^2} + [\varphi''(x) + a(x)\varphi'(x)] \frac{dy}{d\xi} + b(x)y = 0 \qquad (75)$$

(check it up!) in which the expression $x(\xi)$ found from relation (74) should be substituted for x. Equating the coefficient in $\frac{dy}{d\xi}$ to zero we obtain $\varphi(x) = \int_{x_0}^{x} \exp\left(-\int_{x_0}^{\xi} a(s)\,ds\right) d\xi$, and equation (75) takes the form

$$\frac{d^2y}{d\xi^2} + b(x(\xi))\left[\exp\left(2\int_{x_0}^{x(\xi)} a(s)\,ds\right)\right] y = 0$$

Equation (73) and the last equation being of a particular form

$$y'' + p(x)y = 0 \qquad (76)$$

we shall limit ourselves in this section to studying equations (70). The coefficient $p(x)$ and the sought-for solution $y(x)$ will be assumed to be real.

Let us begin with the simplest equation

$$y'' + py = 0 \qquad (77)$$

with constant coefficient p. There is an essential distinction in the behaviour of the solutions of this equation for $p > 0$ and $p < 0$. If $p > 0$ we obtain the general solution

$$y = C_1 \cos \sqrt{p}\,x + C_2 \sin \sqrt{p}\,x = M \sin(\sqrt{p}\,x + \alpha)$$

Every solution of this equation (except, of course, the one identically equal to zero which is of no interest) changes its sign infinitely many times (**oscillates**, as we say), the distance between two neighbouring zeros being $\frac{\pi}{\sqrt{p}}$ and thus decreasing as p grows. For any fixed p the zeros of two linearly independent (not proportional) solutions alternate, that is between any two subsequent zeros of one solution there is exactly one zero of the other.

The situation is quite different when $p < 0$ in equation (77). In this case none of the solutions changes its sign more than once (why?), and therefore we speak of them as **nonoscillating**.

It turns out that the situation remains similar in the case of the general equation (76) although the exact analytic expression of the solution is not available for an arbitrary variable coefficient $p(x)$. J.C.F. Sturm (1803-1855), a French mathematician, estab-

lished in 1836 the following result: *consider equation* (76) *and an auxiliary equation*

$$z''+q(x)\,z=0 \tag{78}$$

of the same type. If the coefficients $p(x)$ *and* $q(x)$ *satisfy the inequality* $p(x)\geqslant q(x)$ *then between any two zeros of every (nonzero) solution of equation* (78) *there is at least one zero of any solution of equation* (76) (this proposition is known as the **Sturm comparison theorem**).

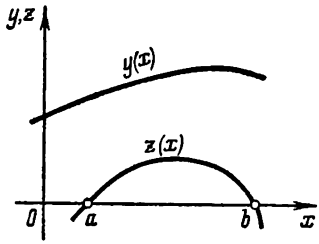


Fig. 122

The theorem is proved by contradiction. Let a and b be two neighbouring zeros of a solution $z(x)$ of equation (78). Suppose that there exists a solution $y(x)$ of equation (76) which has no zero between the points a and b and that both solutions are positive for $a<x<b$ (Fig. 122) (the latter assumption does not violate generality since any solution can be multiplied by -1). Integrating the equality

$$(yz'-y'z)'=yz''-y''z=[p(x)-q(x)]\,yz$$

from a to b we obtain

$$y(b)\,z'(b)-y(a)\,z'(a)=\int_a^b [p(x)-q(x)]\,y(x)\,z(x)\,dx \tag{79}$$

Now the comparison of the signs of the members of equality (79) leads to a contradiction (why?), which completes the proof of the theorem.

The Sturm theorem proved above also includes the case $p(x)\equiv q(x)$ when two solutions of the same equation are compared. (In this case, when analysing equality (79) we should take into account that if $z(a)=0$ then $z'(a)\neq 0$ because, if otherwise, by the uniqueness of the solution $z(x)$ of equation (78) satisfying the given initial conditions $z(a)=z'(a)=0$, we should obtain $z(x)\equiv 0$.) In other words, we arrive at the **Sturm separation theorem:** *given any two linearly independent solutions of equation* (76), *their zeros alternate.* Finally, note that *if two solutions of equation* (76) *possess a common zero they are proportional, that is linearly dependent* (why?).

Now, applying the Sturm comparison theorem to equations (76) and (77) we obtain the following corollaries (let the reader prove them):

1. *If* $p(x)\leqslant 0$, *none of the solutions of equation* (76) *changes its sign more than once, that is there are no oscillating solutions*

2. *If* $p(x) \geqslant p_1 = \text{const} > 0$ *there is no solution of equation* (76) *retaining its sign on an interval whose length exceeds* $\frac{\pi}{\sqrt{p_1}}$.

3. *If* $p(x) \leqslant p_2$ *where* $p_2 = \text{const} > 0$ *the distance between two zeros of any (nonzero) solution of equation* (76) *cannot be less than* $\frac{\pi}{\sqrt{p_2}}$.

4. *If* $p(x) \xrightarrow[x \to \infty]{} p > 0$ *then every (nonzero) solution of equation* (76) *oscillates indefinitely for* $x \to \infty$, *the distance between two adjacent zeros tending to* $\frac{\pi}{\sqrt{p}}$.

Let us take as an example equation (61). The change of type (71) is in this case written as $y = \frac{z}{\sqrt{x}}$ and leads to the equation of the type of (73):

$$z'' + \left(1 - \frac{p^2 - \frac{1}{4}}{x^2}\right) z = 0$$

By Corollary 4, we now conclude that Bessel's function of any order oscillates indefinitely for $x \to \infty$, the distance between two neighbouring zeros tending to π. The same result can be derived from the asymptotic expressions (II.4.28).

A more accurate test for distinguishing oscillating solutions of equation (76) considered on an interval $x_0 \leqslant x < \infty$ can be obtained if we compare it with the equation

$$z'' + \frac{k}{x^2} z = 0 \qquad (0 < x < \infty)$$

The direct integration of this equation (we leave this to the reader; e.g. see [107], Sec. XV.19) implies that for $k > \frac{1}{4}$ all the solutions have infinitely many zeros as $x \to \infty$ while for $k \leqslant \frac{1}{4}$ every solution has no more than one zero. Consequently, by the Sturm theorem, if $\lim\limits_{x \to \infty} x^2 p(x) > \frac{1}{4}$ all the solutions of equation (76) oscillate indefinitely for $x \to \infty$ and if $\lim\limits_{x \to \infty} x^2 p(x) < \frac{1}{4}$ every nonzero solution may only have a finite number of zeros.

In conclusion we note that if we put $y' = z$ in the general equation (70) this results in the system of the first order

$$\left.\begin{aligned} y' &= z \\ z' &= -b(x)\, y - a(x)\, z \end{aligned}\right\}$$

According to Sec. 1, the Wronskian of such a system is expressed as

$$W(x)=\begin{vmatrix} y_1 & y_2 \\ z_1 & z_2 \end{vmatrix}=y_1y_2'-y_2y_1'$$

where the subscripts are the numbers of the solutions. In this case formula (9) takes the form

$$W(x)=W(x_0)\exp\left[-\int_{x_0}^{x}a(\zeta)\,d\zeta\right]$$

and, in particular, for equation (76) we obtain $W(x)=\text{const}$.

11. Systems Involving Parameters. Let us consider the system

$$\dot{\mathbf{y}}=\tilde{\mathbf{A}}(t,\alpha)\,\mathbf{y} \tag{80}$$

containing a parameter α. Its solution satisfying some fixed initial conditions also depends on α: $\mathbf{y}=\mathbf{y}(t,\alpha)$. If the coefficients of the equation are continuous in α the solution is a continuous function of α as well; this is evident from the geometric meaning of a solution of a differential equation and can be proved rigorously on the basis of the lemma given in Sec. 9 (how?). A more complicated situation arises when the dependence of the coefficients on α is discontinuous, for instance, when there appear poles (see Sec. II.3.2). Below we present some considerations related to this case.

Suppose that there is a simple pole of $\tilde{\mathbf{A}}(t,\alpha)$ whose location is independent of t. In particular, this means that if this pole is at a point α_0 the system does not involve coefficients of the form $(t-\alpha)^{-1}$. Making the transformation $\alpha\longrightarrow\alpha_0+\alpha$ we can always reduce the system to the case $\alpha_0=0$, and therefore in what follows we consider a pole at the point $\alpha=0$. Multiplying both sides of system (80) by α we then come to a system of the form

$$\alpha\dot{\mathbf{y}}=\mathbf{A}(t,\alpha)\,\mathbf{y} \tag{81}$$

whose coefficients entering into the right-hand side are continuous in α for small values of $|\alpha|$ and even can be expanded into Taylor's series:

$$\mathbf{A}(t,\alpha)=\sum_{k=0}^{\infty}\mathbf{A}_k(t)\,\alpha^k \tag{82}$$

We shall investigate the structure of the fundamental matrix of system (81) for small $|\alpha|$. To this end we begin with the particular case $n=1$, that is take a scalar equation of type (81):

$$\alpha\dot{y}=y\sum_{k=0}^{\infty}A_k(t)\,\alpha^k,\quad\text{that is}\quad\frac{dy}{y}=\frac{1}{\alpha}\sum_{k=0}^{\infty}A_k(t)\,\alpha^k\,dt$$

It is readily integrated, and its general solution is

$$y = C \exp \int_{t_0}^{t} \sum_{k=0}^{\infty} A_k(t)\, dt \alpha^{k-1} = C \exp \sum_{k=0}^{\infty} \int_{t_0}^{t} A_{k+1}(t)\, dt \alpha^k \cdot \exp \int_{t_0}^{t} \frac{A_0(t)}{\alpha} dt$$

The first exponential expression on the right-hand side can be expanded in nonnegative powers of α, which finally yields

$$y = C \sum_{k=0}^{\infty} B_k(t)\, \alpha^k \cdot \exp \int_{t_0}^{t} \frac{A_0(t)}{\alpha} dt \tag{83}$$

where C is an arbitrary constant and the coefficients $B_k(t)$ can be found with the help of the method of undetermined coefficients by substituting expression (83) into (81) and equating the coefficients in like powers of α (let the reader determine $B_0(t)$ and $B_1(t)$; this involves ambiguity which is accounted for by the fact that C in formula (83) can be replaced by any power expansion corresponding to an expression of the form $C(\alpha)$. This of course affects the coefficients B_k; to avoid the ambiguity we can introduce a normalization condition, for instance, set certain values of these coefficients for $t = t_0$, say $B_0(t_0) = 1$, $B_1(t_0) = B_2(t_0) = \ldots = 0$).

Now we can formulate (without proof) the result of a similar investigation of general systems (81) for which the structure of the general solution turns out to be analogous to (83). (For various results related to asymptotic expansions with respect to independent variables and parameters for solutions of initial- and boundary-value problems for linear and nonlinear differential equations we refer the reader to [137].)

Thus, let us assume that in a neighbourhood of a point $t = t_0$ the coefficient matrix admits of expansion (82), the series on the right-hand side being convergent in the ordinary sense or asymptotically (Sec. II.4.1) for $\alpha \to 0$. We shall suppose that all the eigenvalues of the matrix $\mathbf{A}_0(t_0)$ are distinct. Let $\mu_1(t)$, $\mu_2(t)$, ..., $\mu_n(t)$ be the eigenvalues of the matrix $\mathbf{A}_0(t)$ for the values of t close to t_0. Then system (81) possesses n linearly independent solutions of the form

$$\mathbf{y}_j(t) = \sum_{k=0}^{\infty} \mathbf{B}_{kj}(t)\, \alpha^k \exp \int_{t_0}^{t} \frac{\mu_j(t)}{\alpha} dt \qquad (j = 1, 2, \ldots, n) \tag{84}$$

where the coefficients $\mathbf{B}_k(t)$ can be found with the aid of the method of undetermined coefficients. To this end we substitute (84) into (81) and thus arrive at the equalities (check it up!)

$$\begin{gathered} \mathbf{A}_0(t)\, \mathbf{B}_{0j}(t) - \mu_j(t)\, \mathbf{B}_{0j}(t) = \mathbf{0} \\ \mathbf{A}_0(t)\, \mathbf{B}_{1j}(t) - \mu_j(t)\, \mathbf{B}_{1j}(t) = \mathbf{B}'_{0j}(t) - \mathbf{A}_1(t)\, \mathbf{B}_{0j}(t), \ldots \end{gathered} \tag{85}$$

From the first equality we find $\mathbf{B}_{0j}(t) = \varphi_{0j}(t)\,\mathbf{a}_j(t)$ where $\mathbf{a}_j(t)$ is an eigenvector of the matrix $\mathbf{A}_0(t)$ which is normalized in a certain way and $\varphi_{0j}(t)$ is a scalar function to be determined. Substituting this expression of $\mathbf{B}_{0j}$ into the second equality (85) we obtain

$$\mathbf{A}_0(t)\,\mathbf{B}_{1j}(t) - \mu_j(t)\,\mathbf{B}_{1j}(t) = \\ = \varphi'_{0j}(t)\,\mathbf{a}_j(t) + \varphi_{0j}(t)\,\mathbf{a}'_j(t) + \varphi_{0j}(t)\,\mathbf{A}_1(t)\,\mathbf{a}_j(t) \qquad (86)$$

For this equation to be solvable with respect to $\mathbf{B}_{1j}$ it is necessary and sufficient (see Sec. IV.1.4) that its right member be orthogonal to the eigenvector $\mathbf{b}_j(t)$ of the matrix $\mathbf{A}_0^*(t)$ belonging to the eigenvalue $\mu_j^*(t)$. This orthogonality condition yields a first-order linear differential equation for $\varphi_{0j}(t)$. Solving this equation we determine this function to within an arbitrary constant factor. Now, by virtue of (86), we obtain $\mathbf{B}_{1j}(t) = \tilde{\mathbf{B}}_{1j}(t) + \varphi_{1j}(t)\,\mathbf{a}_j(t)$ where $\tilde{\mathbf{B}}_{1j}(t)$ is a completely determined vector function of t while $\varphi_{1j}(t)$ is a scalar function yet unknown. From the solvability condition for the third equation (85) we now derive a differential equation for $\varphi_{1j}(t)$, and so on.

It turns out that in the general case we can only guarantee the asymptotic convergence of series (84), which does not affect practical significance of the results because we always restrict ourselves to a small number of terms of the series. The expansion obtained applies to any finite and closed interval of the t-axis containing the point t_0 on condition that for this interval all the eigenvalues of the matrix $\mathbf{A}_0(t)$ are distinct. At the same time, this construction has an essentially local character with respect to α, and the range of the values of α to which it applies is usually determined by means of numerical integration of some simpler equations. If the dependence of the coefficients on α involves several poles it is rather difficult to establish the relationship between the asymptotic solutions constructed in the vicinity of different poles although this sometimes can be achieved with the aid of numerical integration or some theoretical considerations, for instance, on the basis of properties common to two or more poles.

As an example, consider the equation

$$\alpha^2 \ddot{y} + [Q(t)]^2 y = 0 \qquad (87)$$

which, after the substitution $y = y_1$, $\alpha\dot{y} = y_2$, reduces to the system

$$\alpha \frac{d}{dt}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -[Q(t)]^2 & 0 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

Taking solution (84) for $j = 0$ and replacing the sum on the right-

hand side by the expression $\exp \sum_{k=0}^{\infty} C_k(t)\,\alpha^k$ (the justification of this technique lies in the ultimate results of the calculations given below) we represent the solution of equation (87) in the form

$$y=\exp\left[\pm\frac{i}{\alpha}\int_{t_0}^{t} Q(t)\,dt+\sum_{k=0}^{\infty} C_k(t)\,\alpha^k\right] \tag{88}$$

To determine the coefficients $C_k(t)$, we substitute (88) into (87). Cancelling by the exponential function and equating the coefficients in the same powers of α we obtain the system of equalities (check it up!)

$$\pm i\dot{Q}\pm 2iQ\dot{C}_0=0,\quad \ddot{C}_0\pm 2iQ\dot{C}_1+\dot{C}_0^2=0,\ \ldots$$

It follows that

$$C_0=-\frac{1}{2}\ln Q(t),\quad C_1=\pm\frac{i}{8}\int_{t_0}^{t}\frac{3\dot{Q}^2-2Q\ddot{Q}}{Q^3}\,dt,\ \ldots$$

Substituting these expressions into (88) we thus determine the sought-for solution of equation (87) up to the terms of the order of α inclusive:

$$y=\frac{1}{\sqrt{Q(t)}}\exp\left[\pm i\int_{t_0}^{t}\left(\frac{Q}{\alpha}+\frac{3\dot{Q}^2-2Q\ddot{Q}}{8Q^3}\,\alpha\right)dt\right] \tag{89}$$

Practically we usually limit ourselves to the foregoing approximation

$$y=\frac{1}{\sqrt{Q(t)}}\exp\left[\pm\frac{i}{\alpha}\int_{t_0}^{t} Q(\tau)\,d\tau\right] \tag{90}$$

The method for constructing solutions of equation (87) which leads to formulas (90) and (89) was independently developed in 1926 by the physicists G. Wentzel, L. Brillouin and H. A. Kramers in connection with some problems of quantum mechanics. Therefore, it is sometimes referred to as the **WBK-method** (and is also known as the **phase integral method**). It is dealt with in [40] and [56] where particular attention is paid to the investigation of the important case when the coefficient $Q(t)$ passes through its zero value (see Sec. 12), for which it is advisable to consider equation (87) for complex values of t.

If the point $\alpha=0$ is a pole of order m ($m\geqslant 1$) for the coefficients of system (80) we take, instead of (81), the equation $\alpha^m\dot{y}=A(t,\alpha)y$ and apply an analogous construction in which the expression

$\int_{t_0}^{t} \frac{\mu_j(t)}{\alpha} dt$ entering into formula (84) should be replaced by

$$\int_{t_0}^{t} \frac{\mu_j(t)}{\alpha^m} dt + \frac{Q_{1,j}(t)}{\alpha^{m-1}} + \ldots + \frac{Q_{m-1,j}(t)}{\alpha}$$

where the coefficients $Q_{kj}(t)$ and $\mathbf{B}_{kj}(t)$ are simultaneously found by means of the method of undetermined coefficients.

12. Turning Points. A point $t = t_0$ is termed a **turning point** of system (81) if the matrix $\mathbf{A}_0(t_0)$ possesses multiple eigenvalues. The simplest example of this kind is the equation

$$\alpha^2 y'' - xy = 0 \tag{91}$$

(here we denote the independent variable by x and mark the differentiation with respect to it by a prime) which is transformed to the system

$$\left.\begin{aligned} \alpha y_1' &= y_2 \\ \alpha y_2' &= x y_1 \end{aligned}\right\}$$

with the coefficient matrix $\mathbf{A}_0(x) = \begin{pmatrix} 0 & 1 \\ x & 0 \end{pmatrix}$ where $y = y_1$ and $\alpha y' = y_2$. The eigenvalues of the matrix being $\mu_{1,2} = \pm \sqrt{x}$, the system has a turning point for $x = 0$.

The change of the independent variable $x = \alpha^{\frac{2}{3}} s$ transforms (91) into **Airy's equation**

$$\frac{d^2y}{ds^2} - sy = 0 \tag{92}$$

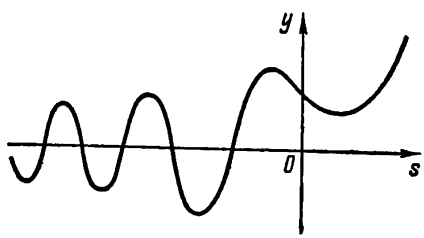


Fig. 123

According to Sec. 10, the behaviour of the solutions to this equation is essentially different for $s > 0$ and for $s < 0$ (see Fig. 123): for $s > 0$ solutions do not oscillate while for $s < 0$ they do, the frequency of the oscillations growing as $s \to -\infty$.

To solve equation (92) we can apply the method of contour integrals in the complex plane similar to the method of Laplace's transformation (§ III.2). We construct the solution in the form of the integral

$$y(s) = \int_{(\Gamma)} e^{sz} v(z)\, dz \tag{93}$$

taken over a contour (Γ) which, together with the function $v(z)$, should be chosen in an appropriate way. The substitution of (93)

into (92) and the integration by parts in the second term lead to the equation

$$z^2v(z)+v'(z)=0$$

provided that the expression $e^{sz}v(z)|_{(\Gamma)}$ is equal to zero. The last equation possesses the apparent solution $v=\exp\left(-\frac{z^3}{3}\right)$. Hence,

$$y(s)=\int\limits_{(\Gamma)}\exp\left(sz-\frac{z^3}{3}\right)dz \qquad (94)$$

(in the general case the contour integral method is applicable to linear differential equations whose coefficients are arbitrary polynomials of degree not higher than the first; see the last example in Sec. III.2.2).

It now becomes evident that we can take as (Γ) any contour whose two branches recede to infinity along two directions for which $\operatorname{Re}(z^3)>0$. These directions lie within the sector

$$-\frac{\pi}{6}<\arg z<\frac{\pi}{6},\quad \frac{\pi}{2}<\arg z<\frac{5}{6}\pi \text{ or } -\frac{5}{6}\pi<\arg z<-\frac{\pi}{2} \qquad (95)$$

The value of integral (94) is of course independent of the particular choice of the contour and is solely determined by the choice of the sectors of type (95) within which the branches of the contour approach infinity. In particular, if the contour of integration passes from the third sector to the second one integral (94) divided by $2\pi i$ is designated as $\operatorname{Ai}(s)$ and called **Airy's function of the first kind**. For real s this function assumes real values and tends to zero for $s\to\infty$ (why?). Applying the saddle-point method (see Sec. II.4.5) it is not difficult to derive the asymptotic representation

$$\operatorname{Ai}(s)\sim\frac{1}{2\sqrt{\pi}}s^{-\frac{1}{4}}\exp\left(-\frac{2}{3}s^{\frac{3}{2}}\right)\qquad(s\to\infty)$$

which is left to the reader. Deforming continuously the contour of integration in (94) in such a way that it is transformed into the imaginary axis we can also deduce the formula

$$\operatorname{Ai}(s)=\frac{1}{\pi}\int\limits_0^\infty\cos\left(\frac{1}{3}t^3+st\right)dt.$$

The other linearly independent solution of equation (92) (denoted by $\operatorname{Bi}(s)$) tends to infinity for $s\to\infty$ as is shown in Fig. 123.

Integral (94) is defined not only for real values of s but also for imaginary ones and is an entire analytic function of the complex variable s. Expanding the integrand in powers of s one can obtain the following expression for Airy's function in terms of

Bessel's functions:

$$\mathrm{Ai}(s) = \frac{1}{3}\sqrt{-s}\left[J_{-\frac{1}{3}}\left(\frac{2}{3}\sqrt{-s^3}\right)+J_{\frac{1}{3}}\left(\frac{2}{3}\sqrt{-s^3}\right)\right]$$

Here all the branches of the functions involved are chosen in such a way that they assume real values for $s=-1$ and are continued in a unique manner to the whole s-plane.

Turning again to equation (91) we obtain its general solution

$$y = C_1\mathrm{Ai}\left(\alpha^{-\frac{2}{3}}x\right)+C_2\mathrm{Bi}\left(\alpha^{-\frac{2}{3}}x\right) \tag{96}$$

We see that the behaviour of the solution is essentially different for $x>0$ and for $x<0$ when $\alpha\to\infty$.

Let us consider the general equation of type (81) with $n=2$ possessing a turning point for $t=0$. This means that the matrix $\mathbf{A}(0, 0)$ must have a two-fold characteristic root (eigenvalue).

The transformation $\mathbf{y}\to e^{\frac{p}{\alpha}t}\mathbf{y}$ diminishes the eigenvalue by p, and therefore, without loss of generality, we can assume that this eigenvalue is equal to zero. Then the determinant $\det \mathbf{A}(t, 0)$ has a zero at the point $t=0$, and we shall assume that this zero is simple. It turns out (see [137]) that under this hypothesis it is possible to make changes of the independent variable and of the unknown functions of the form

$$\mathbf{y} = \exp\left(\frac{1}{2\alpha}\int_0^t \mathrm{tr}\,\mathbf{A}(\tau, 0)\,d\tau\right)\mathbf{P}(t)\,\mathbf{u}, \quad t=\varphi(s) \tag{97}$$

with appropriately chosen $\mathbf{P}(t)$ and $\varphi(s)$ such that system (81) is transformed to an analogous system of the form

$$\alpha\dot{\mathbf{u}} = \mathbf{B}(s, \alpha)\,\mathbf{u} \tag{98}$$

with $\mathbf{B}(s, 0)=\begin{pmatrix}0 & 1\\ s & 0\end{pmatrix}$. Now we can make one more change

$$\mathbf{u} = \sum_{k=0}^{\infty} Q_k(s)\,\alpha^k\mathbf{v} \tag{99}$$

transforming system (98) to

$$\alpha\dot{\mathbf{v}} = \begin{pmatrix}0 & 1\\ s & 0\end{pmatrix}\mathbf{v} \tag{100}$$

The coefficients of transformation (99) as well as the functions $\mathbf{P}(t)$ and $\varphi(s)$ in (97) can be found with the aid of the method of undetermined coefficients.

System (100) being equivalent to equation (91) and all the transformations performed being nondegenerate in the vicinity of the

point $t=0$, $\alpha=0$, in the general case as well the solution is expressible, in the vicinity of the turning point, in terms of Airy's functions (see (96)), that is has singularities of the same type as in example (91).

§ 2. AUTONOMOUS SYSTEMS

As is known (e.g. see [107], Sec. XV.12), a system of differential equations of the form

$$\left.\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2, \ldots, x_n) \\ \dot{x}_2 &= f_2(x_1, x_2, \ldots, x_n) \\ &\cdots\cdots\cdots\cdots \\ \dot{x}_n &= f_n(x_1, x_2, \ldots, x_n) \end{aligned}\right\} \tag{1}$$

whose right members do not involve the independent variable t is called **autonomous.** The general approach to such systems is based on some geometric considerations widely used in modern mathematics. Here we shall briefly discuss them. For more detail we refer the reader to [4] and [82].

In what follows all the quantities involved are supposed to be real.

1. General Concepts. It is convenient to consider the solutions of system (1) as representing curves in the space of the variables $x_1, x_2, \ldots, x_n$ called the **phase space** of the system. Its points will be designated by a single letter, for instance, $x=(x_1, x_2, \ldots, x_n)$. Moreover, since the coordinates, as a rule, will not be used separately we shall agree that a letter supplied by a subscript designates a point of the phase space but not a coordinate unless otherwise stated. The radius vector of a point x will be denoted as $\mathbf{x}$. (Thus, under this convention, the symbol x does not designate the modulus of the vector $\mathbf{x}$!)

Multiplying each equation (1) by the unit vector of the corresponding coordinate axis and adding together the results termwise we arrive at the shortened form of the same system:

$$\dot{\mathbf{x}} = \mathbf{f}(x) \tag{2}$$

Thus, we can say that there is a field of velocities defined in the phase space, and it is natural to interpret it as a stationary flow, every concrete solution $x=x(t)$ describing the law of motion of a particle of the flow.

We shall suppose that the right-hand sides of system (2) satisfy the well-known sufficient conditions guaranteeing the existence and the uniqueness of the solution of Cauchy's problem $x(0)=x_0$ (e.g. see [107], Sec. XV.12). This solution will be denoted by $x=x(t, x_0)$. According to what has been said, $x(t, x_0)$ is a continuous function

satisfying the conditions

$$x(0, x_0) = x_0, \quad x(t, x(t_1, x_0)) = x(t + t_1, x_0) \tag{3}$$

(Let the reader examine the second equality expressing the stationarity of the flow.)

In modern mathematics we speak of a *flow* defined in a space if there is a one-parameter family $x = x(t, x_0)$ (the parameter t is usually interpreted as time) of mappings of the space into itself, satisfying conditions (3), irrespective of the concrete nature of the quantities involved, that is irrespective of any originally given differential equations.

If $x(t)$ is a solution of system (2) the oriented curve described by the moving point $x = x(t)$ in the phase space, as t increases, is called a **(phase) trajectory** of the system. For any t_0 the function $x(t - t_0)$ is also a solution of the system specifying the same trajectory traced by the moving point with time lag t_0.

For the sake of simplicity, we shall suppose that system (2) is defined in the whole space E_n and that every solution exists for $-\infty < t < \infty$. It is sometimes necessary to consider a system which is only defined in a domain of E_n and also the cases when some trajectories recede to infinity for finite t, which does not involve essential changes in the general scheme of investigation.

There can also be a trajectory consisting of a single point. Such a trajectory is referred to as a **rest point** (or a **stationary point** or an **equilibrium solution**). For x_0 to be a rest point it is necessary and sufficient that

$$\mathbf{f}(x_0) = \mathbf{0} \tag{4}$$

(why?).

Since the flow is stationary, a trajectory along which the moving point passes twice through the same point is a closed line (referred to as a **cycle**), the corresponding solution $x(t)$ being periodic. Thus, we deal with the following three basic types of trajectories: nonclosed, closed (cycles) and rest points. Every point of the phase space belongs exactly to one trajectory.

2. Limiting Behaviour of Solutions. To characterize the behaviour of a nonclosed trajectory we introduce the notion of *limit sets* of the trajectory. Let $x(t)$ be a solution of system (2) specifying a trajectory (l). The collection of the limits of all the possible convergent sequences of the form $\{x(t_i)\}$ where $t_i \longrightarrow \infty$ is called the **ω-limit set** of the trajectory; similarly, if $t_i \longrightarrow -\infty$ we speak of the **α-limit set**. For instance, in Fig. 124 we see a trajectory whose α-limit set consists of a single point a while its ω-limit set is a cycle (L). The figure shows that the sequence $x(1), x(8), x(16), \ldots$ converges to the point $p \in (L)$, and it is clear that such a sequence can also be constructed for any point of the cycle (L).

We shall enumerate some properties of limit sets (proved in mathematical courses dealing with the theory of autonomous systems) which are rather evident. For definiteness, we shall consider ω-limit sets. (Note that if the time t is reversed the roles of the α- and ω-limit sets are interchanged.)

1. *If the ω-limit set of a trajectory is an empty (void) set*, that is contains no points, *the trajectory recedes to infinity for* $t \to \infty$ (in the sense that $|\mathbf{x}(t)| \xrightarrow[t \to \infty]{} \infty$). (If the space E_n is completed by adding the point at infinity to it, as was done for the complex plane in Sec. II.1.7, we can say that the ω-limit set of such a trajectory is the point at infinity.)

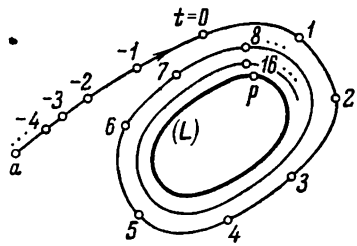


Fig. 124

2. *If the ω-limit set consists of a single point* **a**, *this is a rest point, and the trajectory enters it for* $t \to \infty$, that is tends to it as $t \to \infty$: $|\mathbf{x}(t) - \mathbf{a}| \xrightarrow[t \to \infty]{} 0$. (Taking into account the possibility of reversing time t we see that the point a shown in Fig. 124 is a rest point.) It should be stressed that a trajectory cannot arrive at a rest point at a finite time moment because, if otherwise, we obtain a contradiction since two different trajectories cannot have a point in common. When we say that a trajectory (l) enters a rest point a for $t \to \infty$ we do not attribute the rest point itself to the trajectory.

3. *An ω-limit set is always closed* (as a point set, that is contains all its limit points).

4. *If* $(\Omega_{(l)})$ *is an ω-limit set it consists of entire trajectories:* if $p \in (\Omega_{(l)})$ the entire trajectory passing through the point p belongs to $(\Omega_{(l)})$.

5. *If an ω-limit set is bounded it is connected.* (If it is not bounded the addition to it of the point at infinity included into the space makes it connected; for example, see Fig. 125 where the ω-limit set consists of two straight lines.)

It is sometimes possible to specify regions in the space filled with trajectories possessing the same asymptotic behaviour, that is having the same ω-limit set. These regions are separated by $(n-1)$-dimensional surfaces called **separatrices**. The limiting behaviour of trajectories on a separatrix is usually distinct from that of the trajectories both lying on one side of it and on the other side. In Fig. 126 we see a possible location of three rest points and one separatrix (l).

3. Rest Points in the Plane. Linear Systems. Let us take system (2) defined in the xy-plane:

$$\dot{x} = P(x, y), \quad \dot{y} = Q(x, y) \tag{5}$$

It follows that the phase trajectories of (5) are described by the differential equation

$$Q(x, y)\,dx - P(x, y)\,dy = 0 \tag{6}$$

and therefore can be constructed approximately with the help of isoclines (e.g. see [107], Sec. XV.3). In this construction it is advisable to draw a sketch of the family of the trajectories in the vicinity of the singular points of equation (6) which are the rest points of system (5) (why?).

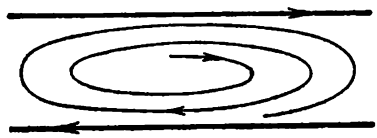


Fig. 125

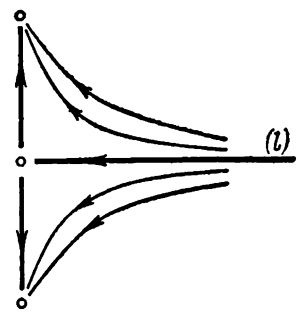


Fig. 126

Expanding the right members $P(x, y)$ and $Q(x, y)$ at a stationary point (x_0, y_0) of system (5) (which is a singular point of equation (6)) in powers of $x-x_0$ and $y-y_0$ and using condition (4) we obtain

$$\dot{x} = a(x-x_0) + b(y-y_0) + \ldots, \quad \dot{y} = c(x-x_0) + d(y-y_0) + \ldots \tag{7}$$

where the dots designate the terms of order higher than the first relative to $x-x_0$ and $y-y_0$ and

$$a = P'_x(x_0, y_0), \quad b = P'_y(x_0, y_0),$$
$$c = Q'_x(x_0, y_0), \quad d = Q'_y(x_0, y_0) \tag{8}$$

Let us suppose that the rest point (x_0, y_0) lies at the origin: $x_0 = y_0 = 0$ (this can always be achieved by the transformation $x \longrightarrow x_0 + x$, $y \longrightarrow y_0 + y$). For small values of $|x|$, $|y|$ it is natural to discard the terms of higher order of smallness and thus replace (7) by the linearized system

$$\dot{x} = ax + by, \quad \dot{y} = cx + dy \tag{9}$$

which can also be written in the vecto: form

$$\dot{\mathbf{r}} = \mathbf{A}\mathbf{r} \tag{10}$$

where

$$\mathbf{r} = \begin{pmatrix} x \\ y \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

We shall impose the condition that the matrix $\mathbf{A}$ is nondegenerate, which reduces to the condition that the origin is an isolated

stationary point of system (9) (why are these conditions equivalent?).

To simplify system (10) we make an affine transformation $\mathbf{r}=\mathbf{Tr}'$ in the xy-plane (e.g. cf. [107], Sec. XI.6). Since $\dot{\mathbf{r}}=\mathbf{Ar}$, the equation of the trajectories after the transformation is $\frac{d}{dt}\mathbf{T}(\mathbf{r}')=\mathbf{A}(\mathbf{Tr}')$, that is

$$\frac{d\mathbf{r}'}{dt}=\mathbf{T}^{-1}\mathbf{ATr}' \tag{11}$$

The course of investigation depends on the roots of the characteristic equation for the matrix $\mathbf{A}$ which is written as

$$\det\begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix}=0$$

that is

$$\lambda^2-(a+d)\lambda+ad-bc=0 \tag{12}$$

We shall consider the following possible cases.

1. *Let the roots* λ_1, λ_2 *of equation* (12) *be real and have opposite signs.* As is known from the theory of matrices (e.g. see [107], Sec. XI.8), in this case it is always possible to choose the matrix $\mathbf{T}$ in such a way that $\mathbf{T}^{-1}\mathbf{AT}=\operatorname{diag}(\lambda_1, \lambda_2)$. Then system (11) takes the form $\frac{dx'}{dt}=\lambda_1 x'$, $\frac{dy'}{dt}=\lambda_2 y'$ whence $x'=C_1e^{\lambda_1 t}$, $y'=C_2e^{\lambda_2 t}$, that is

$$y'=C|x'|^{\frac{\lambda_2}{\lambda_1}} \tag{13}$$

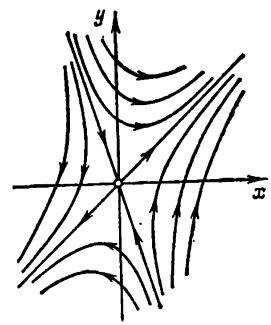


Fig. 127

We thus obtain a family of lines, resembling hyperbolas, with x'- and y'-axes as asymptotes, the latter also being phase trajectories. In the original xy-plane we obtain the transformed family of the trajectories (see Fig. 127). As is known (e.g. see [107], Fig. 290), a singular point of this type is termed a **saddle point**.

The directions along which the two rectilinear trajectories (the asymptotes) pass through the saddle point can be readily found. Indeed, if the equation of such a straight line is $y=kx$, equations (9) imply that

$$k=\frac{c+dk}{a+bk} \tag{14}$$

(check it up!) whence both possible values of k can easily be determined.

To determine the orientation of the phase trajectories, that is to find out in what directions the moving point travels along them,

it suffices to take an arbitrary (nonsingular) point, construct the velocity vector (9) at it and then continuously extend the direction thus determined to the other trajectories.

2. *Let λ_1 and λ_2 be real, distinct and have the same sign.* Then, as in Case 1, we arrive at equality (13) but the structure of the family of trajectories in the vicinity of the rest point differs from the above (see Fig. 128): all the trajectories enter the rest point (or start from it if the arrows in the figure are reversed), each trajectory having a certain direction near the rest point. A stationary point of this type is called a **nodal point**. More precisely, in this case we obtain a node with two directions along which the trajectories tend to or go away from it: all the trajectories, except two ones (*AO* and *BO* in Fig. 128), have the same direction (if we do not distinguish between the opposite directions having the same slope) while the two exceptional trajectories go along another direction.

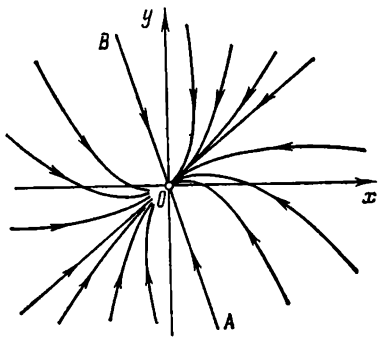


Fig. 128

3. *Let both roots be imaginary but not pure imaginary*: $\lambda_{1,2} = \mu \pm i\nu$, $\mu \neq 0$. Then there is no real transformation matrix **T** reducing the matrix **A** to diagonal form. In this case, as was proved in Sec. IV.3.4, the matrix **A** can be brought to the form $\begin{pmatrix} \mu & \nu \\ -\nu & \mu \end{pmatrix}$, and then system (11) is written as

$$\frac{dx'}{dt} = \mu x' + \nu y', \quad \frac{dy'}{dt} = -\nu x' + \mu y'$$

Passing to polar coordinates in the $x'y'$-plane we obtain (check up the calculations!)

$$\dot{\rho} = \mu\rho, \quad \dot{\varphi} = -\nu \quad \text{whence} \quad \rho = Ce^{-\frac{\mu}{\nu}\varphi}$$

The last equation describes logarithmic spirals, and in the original xy-plane we get the transformed family of spirals (Fig. 129). Such a rest point is called **focal.**

4. *Let the roots be pure imaginary.* Then we similarly obtain $\rho = C$, which means that in the xy-plane there is a family of similar ellipses with common symmetry axes. A point of this kind is termed a **centre** (Fig. 130).

5. *Let there be a two-fold root* λ. Then, according to Sec. IV.3.3, the matrix **A** can be reduced to the form $\begin{pmatrix} \lambda & 0 \\ 1 & \lambda \end{pmatrix}$ or $\begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$. The second subcase only takes place when $a=d$, $b=c=0$ (why?).

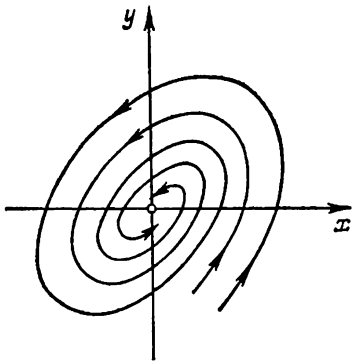


Fig. 129

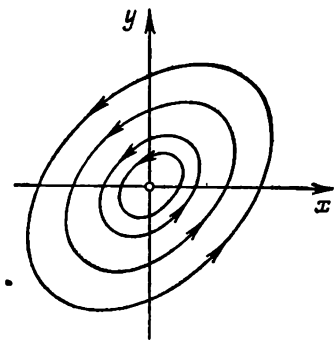


Fig. 130

In the first subcase system (11) assumes the form

$$\frac{dx'}{dt}=\lambda x', \quad \frac{dy'}{dt}=x'+\lambda y'$$

and the integration yields (check up the result!)

$$\lambda y'=x'(\ln|x'|+C)$$

We thus obtain a nodal point with a single direction along which the moving points of the trajectories go in the vicinity of the point (see Fig. 131).

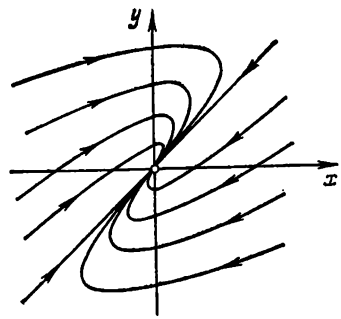


Fig. 131

In the second subcase we receive $y=Cx$, which means that the trajectories are the straight lines passing through the origin. This is also a nodal point (sometimes referred to as a **degenerate node**).

If the coefficients a, b, c and d in system (9) are regarded as arbitrary parameters, the basic cases are the first three since they are specified by certain inequalities connecting the coefficients (for instance, in Case 1 there must be $ad-bc<0$ and $(a+d)^2-4(ad-bc)>0$, that is there are four degrees of freedom in the choice of the coefficients, while in Cases 4 and 5 we have only three degrees of freedom (why?)). Besides, since the inequalities are strict

they remain valid when the coefficients are given sufficiently small variations, which means that the rest points of the first three types are *structurally stable* whereas those belonging to the last two types may change into rest points of other types (which?) under such a variation.

A nodal and a focal point can be stable or unstable depending on whether the motion along the trajectories is directed toward the rest point or from it. Using the basic test for Lyapunov stability (e.g. see [107], Sec. XV.22) and equation (12) it is possible to show that for a nodal or a focal point to be stable it is necessary and sufficient that $a+d<0$ (the proof is left to the reader).

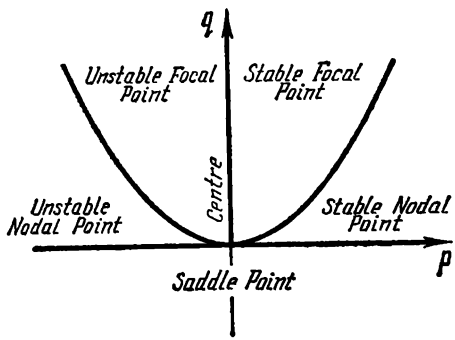


Fig. 132

Denoting the coefficients of the characteristic equation (12) as $p=-(a+d)$ and $q=ad-bc$ we obtain the regions in the p, q-plane (see Fig. 132) corresponding to different types of rest points. The line separating the nodes from the focuses has the equation $q=\frac{p^2}{4}$; its points correspond to the nodes described in Case 5.

4. General Case. Here we return to general system (7). A more thorough analysis carried out in comprehensive mathematical courses shows that the rest point (x_0, y_0) of this system retains the same type as the point (0, 0) of system (9). All the trajectories shown in Figs. 127-131 are of course deformed in the general case but in a small neighbourhood of each rest point the shape of the trajectories and the directions along which they go in the vicinity of the rest point are invariable while for the regions lying far from the rest point the distortion of the trajectories may be considerable. Therefore, it is natural to retain in the nonlinear case the classification of rest points given for linear systems since it is related to the behaviour of the trajectories in small neighbourhoods of the points. We thus can say that, generally speaking, the linearization of a nonlinear system does not change the type of a rest point and, conversely, that the addition of nonlinear terms to a linear system preserves the type. The only exception to this general rule is the case of a centre because the addition of terms of arbitrarily high order of smallness may change a centre into a focus (see Fig. 133). In real nonlinear systems describing natural phenomena this type of a rest point hardly ever occurs unless

there are special circumstances specifying conditions (see Sec. 4.2) under which a centre can appear.

We note that when determining the rest points of a concrete system (5) we have to solve the finite nonlinear equations $P = Q = 0$ but if we are only interested in the type of the rest points for structurally stable cases (Sec. 3) there is no need in high accuracy of the calculations. Indeed, the type of a structurally stable rest point is specified by strict inequalities imposed on the coefficients of the linearized system which are equal to the derivatives of the right members (see (8)). If the derivatives are continuous the inequalities connecting their values at a rest point hold in a small neighbourhood of the rest point as well. The roots of equation (14) determining the directions along which the trajectories go in the vicinity of a rest point are also stable with respect to small variations of the derivatives.

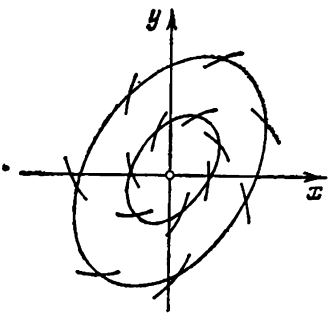


Fig. 133

If among the rest points of a given system there are saddle points then, in order to have a correct judgement upon the location of phase trajectories in the plane, it is expedient to specify the corresponding separatrices. To this end, we can apply numerical integration following the scheme given below. We begin with the determination of the direction of a separatrix passing through a given rest point by applying equation (14) to the linearized system. Then, at a certain distance from the rest point, we draw an axis perpendicular to this direction (Fig. 134) and take its points as initial ones for numerical integration carried out both for increasing and decreasing values of t (Fig. 134 corresponds to the case of backward integration). This makes it possible to specify the course of the trajectories after they leave the vicinity of the saddle point, which enables us to see on what side of the separatrix the initial points lie. After the location of a point belonging to the separatrix has thus been approximately determined the separatrix is continued by means of numerical integration.

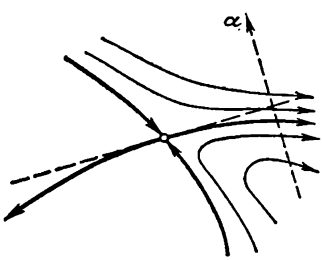


Fig. 134

One must not think that the types of rest points studied above exhaust all the possibilities. If **A** is a degenerate matrix (particularly, equal to zero) the linear approximation is inapplicable, and to investigate the character of the rest point it is necessary

to resort to more sophisticated techniques showing that there exists a vast variety of different types of rest points depending on the structure of the terms of higher order of smallness.

It is sometimes possible to study the structure of a rest point belonging to these more complicated types by passing to polar coordinates ρ and φ: $x=\rho\cos\varphi$, $y=\rho\sin\varphi$. The substitution of these expressions into (5) results in the equations (check up the calculations!)

$$\left.\begin{aligned}\dot{\rho}&=P(\rho\cos\varphi,\ \rho\sin\varphi)\cos\varphi+Q(\rho\cos\varphi,\ \rho\sin\varphi)\sin\varphi\\ \rho\dot{\varphi}&=Q(\rho\cos\varphi,\ \rho\sin\varphi)\cos\varphi-P(\rho\cos\varphi,\ \rho\sin\varphi)\sin\varphi\end{aligned}\right\}$$

whence

$$\frac{d\rho}{d\varphi}=\rho\,\frac{P(\rho\cos\varphi,\ \rho\sin\varphi)\cos\varphi+Q(\rho\cos\varphi,\ \rho\sin\varphi)\sin\varphi}{Q(\rho\cos\varphi,\ \rho\sin\varphi)\cos\varphi-P(\rho\cos\varphi,\ \rho\sin\varphi)\sin\varphi} \tag{15}$$

Suppose that the expansions of the functions P and Q in powers of x and y begin with terms of order m ($m\geqslant 1$), that is

$$P(x, y)=P_m(x, y)+\ldots,\ Q(x, y)=Q_m(x, y)+\ldots$$

where P_m and Q_m are homogeneous polynomials of degree m (from which at least one is not identically equal to zero) and the dots designate the terms of higher (than m) order of smallness. Substituting the expansions into (15) and cancelling we obtain

$$\frac{d\rho}{d\varphi}=\rho\,\frac{A(\varphi)+\ldots}{B(\varphi)+\ldots} \tag{16}$$

where we have put, for brevity,

$$A(\varphi)=P_m(\cos\varphi,\ \sin\varphi)\cos\varphi+Q_m(\cos\varphi,\ \sin\varphi)\sin\varphi$$

and

$$B(\varphi)=Q_m(\cos\varphi,\ \sin\varphi)\cos\varphi-P_m(\cos\varphi,\ \sin\varphi)\sin\varphi$$

(here the dots denote the terms involving ρ). It should be noted that

$$A(\varphi+\pi)\equiv(-1)^{m+1}A(\varphi),\ B(\varphi+\pi)\equiv(-1)^{m+1}B(\varphi)$$

Let us interpret φ and ρ as Cartesian coordinates in an auxiliary plane and consider (16) as a differential equation specifying phase trajectories in this plane. It should be taken into account that the right-hand side of (16) is periodic in φ with period 2π and that we only are interested in small values of ρ. If $B(\varphi)\not\equiv 0$ for $0\leqslant\varphi\leqslant\pi$ we readily see that, since equation (16) possesses the particular solution $\rho(\varphi)\equiv 0$, the family of the integral curves is approximately of the shape shown in Fig. 135. Coming back to the xy-plane we conclude that in this case the origin is a focal

point (in some special cases it may be a centre). For instance, Fig. 135 corresponds to the case of spirals winding about the singular point in the positive direction (why?).

Now suppose that $B(\varphi)$ has zeros but is not identically equal to zero. Then there are singular points of equation (16) on the φ-axis at these zero points. For the sake of simplicity we shall assume that $B(\varphi_0)=0$, $B'(\varphi_0)\neq 0$ and $A(\varphi_0)\neq 0$. Then, according to the above, we can truncate equation (16) and write

$$\frac{d\rho}{d\varphi}=k\frac{\rho}{\varphi-\varphi_0} \quad \text{where} \quad k=\frac{A(\varphi_0)}{B'(\varphi_0)}$$

Thus, if $k>0$ there is a node at the point φ_0 and if $k<0$ a saddle point. In this way we can test separately each zero of the function $B(\varphi)$. (Note that $B(\varphi)=0$ is an algebraic equation of the $(m+1)$th degree relative to $\tan\varphi$.)

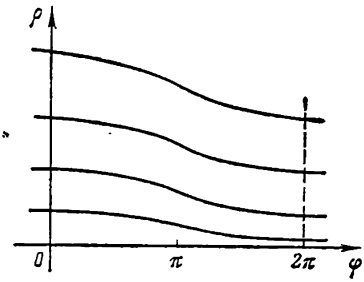


Fig. 135

After the structure of the trajectories in the $\varphi\rho$-plane has been elucidated we return to the xy-plane. For example, suppose that on the φ-axis there are two neighbouring singular points of equation (16): a nodal point $\varphi=\varphi_1$ and a saddle point $\varphi=\varphi_2$ (see

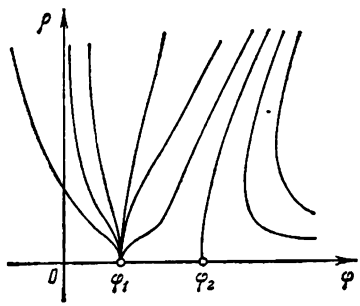


Fig. 136

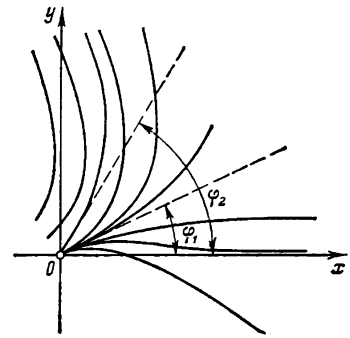


Fig. 137

Fig. 136). Then in a sector in the xy-plane with vertex at the origin the trajectories are located as depicted in Fig. 137: there is a family of trajectories approaching the origin in the direction with polar angle φ_1 while in the direction with polar angle φ_2 only one trajectory goes. This is an example of a complete investigation of the structure of the neighbourhood of a rest point of higher order.

If $A(\varphi) \equiv 0$ or $B(\varphi) \equiv 0$, in (15) we should use, respectively, subsequent terms of the expansion of the numerator or of the denominator.

To obtain a full description of phase trajectories in the phase plane it is as well advisable to investigate their behaviour at infinity. To this end we can, for instance, perform an inversion transformation (see Sec. II.1.8) or, assuming that the ratio $\frac{y}{x}$ is bounded, make the transformation $\xi = \frac{1}{x}$, $\eta = \frac{y}{x}$ (find out its meaning!) and also, for bounded $\frac{x}{y}$, use the transformation $\tilde{\xi} = \frac{1}{y}$, $\tilde{\eta} = \frac{x}{y}$. The information obtained in this way must be reformulated in terms of the coordinates x and y. The neighbourhood of the point at infinity is considered analogously for phase spaces of any dimension.

5. Cycles in the Plane. A cycle whose sufficiently small neighbourhood does not contain other cycles (i.e. an **isolated cycle**) can

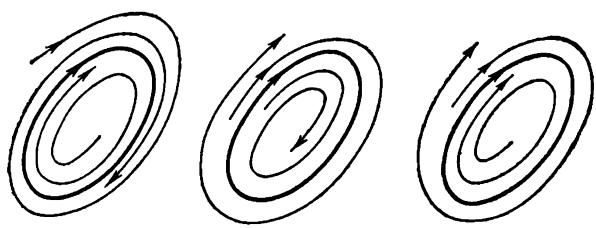


Fig. 138

be **stable, unstable** and, more rarely, **half-stable** (see Fig. 138) depending on the behaviour of the trajectories in its vicinity. For instance, all the trajectories lying close to a stable cycle, when continued indefinitely, wind around the cycle approaching it spirally from the inside or from the outside, that is have this cycle as their ω-limit set (Sec. 2). Generally, a closed trajectory which is an α- or an ω-limit set of a nonclosed trajectory is called a **limit cycle.** Thus, isolated cycles in the plane are limit ones. The case when there is a region in the plane entirely filled with cycles is also possible; such cycles are of course nonisolated.

The motion described by a stable limit cycle is called an **auto-oscillation.** A feature of such a motion is that for small perturbations (more precisely, for the perturbations which do not make the moving point leave the so-called **domain of influence** or **domain of attraction** of the cycle) the perturbed motion approaches the

cycle asymptotically (for a real physical system this means that, after a transient regime, the system returns to the stable cyclic motion).

When determining limit cycles we can use the following theorem which we give without proof (its assertion is rather visual): *if there is a trajectory which, for $t \geqslant t_0$, entirely lies within a finite domain, without rest points in its interior and on the boundary, the trajectory is either closed or spirally approaches a closed trajectory (that is winds around it indefinitely and tends to it)*. For instance, this theorem can be used in the following way. Assume that it is possible to construct a ring-shaped region (see Fig. 139) on whose boundary the trajectories enter the interior of the region (this property can readily be checked by constructing the field of velocities on the boundary) and that the region and its boundary do not contain rest points (which are determined directly from condition (4)). Then, by the theorem, taking a trajectory starting at any point inside the region or on its boundary and continuing it for increasing t by means of numerical integration of system (5) we arrive at a cycle (theoretically, in the limit but, practically, for a sufficiently large finite time t). If the trajectories leave the region on its boundary the same construction can be realized for decreasing values of t. In practical applications it is not necessary to construct such an annular region because it is often sufficient to resort to numerical integration beginning with various points of the domain where a cycle is expected to lie and construct the trajectories thus found on the plotting paper, which, as a rule, enables us to determine the location of the cycle. For this purpose it is also possible to apply graphical integration by using a sufficiently dense net of isoclines (e.g. see [107], Sec. XV.3) or some other procedure of the type described in courses of the theory of vibrations.

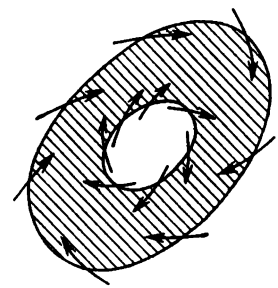


Fig. 139

The following proposition can also be of use: *if the sum $P'_x + Q'_y$ does not change its sign in a simply connected domain (G) and is not identically equal to zero there is no cycle entirely contained in (G)*. In fact, if we suppose that (l) is such a cycle bounding a plane figure (S) within (G), Ostrogradsky's theorem implies that

$$\oint_{(l)} (P\mathbf{i} + Q\mathbf{j})_n \, dl = \iint_{(S)} (P'_x + Q'_y) \, dS$$

The left member is equal to zero while the right one is nonzero (why?) and hence we arrive at a contradiction, which completes

the proof. (Analogously, if in a k-connected domain G the sum $P'_x + Q'_y$ retains the sign and is not identically zero the number of cycles lying in G is less than k.)

For investigating cycles and solving some other problems the following construction is often applied. Let us arbitrarily choose two arcs (l) and (l') which are at no point tangent to the field of velocities corresponding to the given system. Let the arcs be such that every trajectory starting at a point of (l) achieves (l') (Fig. 140). This specifies a point mapping of (l) onto (l'). If α is a coordinate reckoned along (l) and α' is a coordinate along (l') the mapping is described by a function $\alpha'(\alpha)$ which is monotone and continuous.

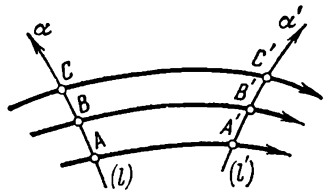


Fig. 140

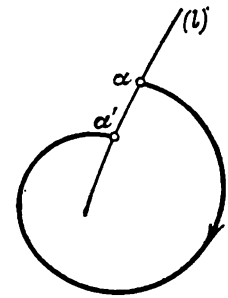


Fig. 141

In particular, if $(l') = (l)$ (see Fig. 141) the function $\alpha'(\alpha)$ **(Poincaré's mapping)** is increasing, and for a value $\alpha = \alpha_0$ to determine a cycle it is necessary and sufficient that the equality

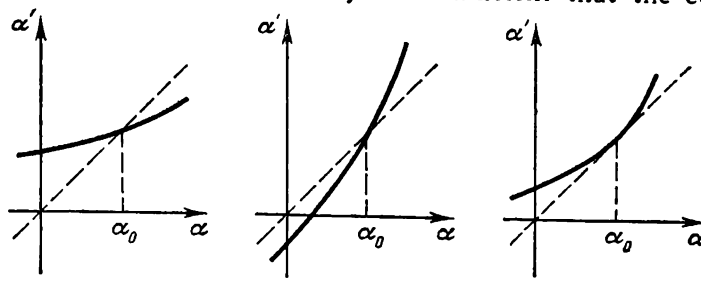


Fig. 142

$\alpha'(\alpha_0) = \alpha_0$ take place or, in other words, that α_0 be a fixed point of the mapping. In Fig. 142*a*, *b* and *c* we see sketches of the graph of the function $\alpha'(\alpha)$ for, respectively, a stable, an unstable and a half-stable limit cycle. By the way, the function $\alpha'(\alpha)$ is usually constructed by means of numerical integration of system (5), and therefore its significance lies mostly in theoretical aspects.

This construction reveals a subtleness of the behaviour of a half-stable limit cycle: if a system (1) be given an arbitrarily small variation such a cycle sometimes disappears (if the variation makes the graph of the function x' intersect upwards) or turns into a pair of closely located cycles of opposite types (if the graph moves downwards). A real significance of a half-stable cycle is that it occurs when a system (1) involves a parameter whose variation makes two cycles merge for a certain value of the parameter.

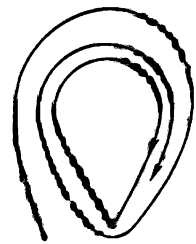


Fig. 143

One must not think that every enclosed trajectory in the plane which is bounded for $t \to +\infty$ must either enter a rest point or wind around a cycle: there can also be a case when, as in Fig. 143, the entire set of such a trajectory consists of several enclosed trajectories and rest points.

6. Index of a Singular Point of a Vector Field. Here we shall consider the notion of the *index of a singular point* (see [70]) which is useful for studying vector fields and, particularly, velocity fields in the plane.

Consider a plane vector field

$$\mathbf{A}(x, y) = P(x, y)\,\mathbf{i} + Q(x, y)\,\mathbf{j}$$

and a closed line (L) in the plane (which is not necessarily a vector line of the field, that is, may not be a trajectory of the

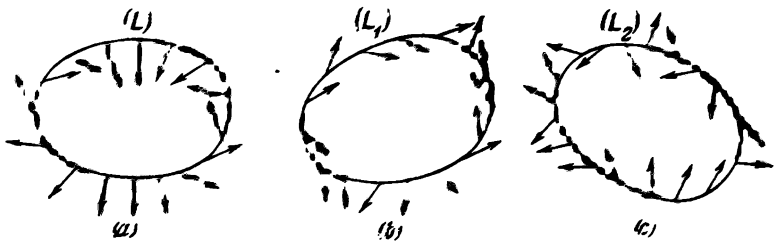


Fig. 144

system of differential equations $\dot{x} = P(x, y)$, $\dot{y} = Q(x, y)$). Let there be no singular points of the field on the contour (L). A singular point is understood here as one for which the vector field vanishes, or has a discontinuity. The index $\gamma(\mathbf{A}, (L))$ of the line (L) relative to the vector field $\mathbf{A}$ is defined as the number of revolutions in the positive sense made by the vector $\mathbf{A}(M)$ when the point M describes the line (L) once in the positive sense (see Fig. 144a, b and c, where respectively the cases $\gamma(\mathbf{A}, (L)) = 2$, $\gamma(\mathbf{A}, (L)) = 0$ and $\gamma(\mathbf{A}, (L)) = -2$ are depicted). The index possesses the following simple properties.

1. *The number* $\gamma[\mathbf{A}, (L)]$ *is always an integer* (positive, negative or zero).

2. *If the contour* (L) *or the field* $\mathbf{A}$ *is continuously deformed in such a way that* (L) *does not cross singular points, the index* $\gamma[\mathbf{A}, (L)]$ *remains invariable.* Indeed, when the contour or the field is given a small variation the index obviously cannot receive a large increment, and if we suppose that it changes, its increment must also be small. On the other hand, the index being an integer, its value cannot gain a small change unless it retains a constant value (why?), which completes the proof.

In particular, it follows that *if there are no singular points of the field inside* (L) *we have* $\gamma[\mathbf{A}, (L)] = 0$ (why?).

3. *If the domain lying inside* (L) *is broken up into* k *parts with boundaries* (L_1), (L_2), ..., (L_k), *then*

$$\gamma[\mathbf{A}, (L)] = \gamma[\mathbf{A}, (L_1)] + \gamma[\mathbf{A}, (L_2)] + \ldots + \gamma[\mathbf{A}, (L_k)]$$

(the proof is left to the reader).

In concrete problems the index can sometimes be determined visually (as in Fig. 144) or with the aid of the formula

$$\gamma[\mathbf{A}, (L)] = \frac{1}{2\pi} \oint\limits_{(L)} d\left(\arctan \frac{Q}{P}\right) = \frac{1}{2\pi} \oint\limits_{(L)} \frac{1}{P^2+Q^2}(P\,dQ - Q\,dP) \quad (17)$$

Now consider an isolated singular point M_0 of the field $\mathbf{A}$. Let (L) be a contour containing the point M_0 in its interior and no other singular points. Such a contour can be chosen in different ways but, according to Property 2, the value of $\gamma[\mathbf{A}, (L)]$ is the same for all such contours. It is referred to as the **index** $\gamma(M_0)$ **of the singular point** M_0 **of the field A.** For instance, from Figs. 127-131 we see that for a plane vector field of velocities the index of a saddle point is equal to -1 while for a nodal point, focal point and centre it is equal to 1. (Let the reader check up these properties.)

Now suppose that a given contour (L) contains inside it singular points M_1, M_2, ..., M_k. Divide the plane domain lying inside the contour (L) into parts in such a way that each of them contains exactly one singular point in its interior. Then the application of Property 3 leads to the formula

$$\gamma[\mathbf{A}, (L)] = \sum_j \gamma(M_j) \quad (18)$$

where the sum is extended over all the singular points lying inside (L). This formula resembles Cauchy's residue theorem (see formula (II.3.9)), and there is in fact an interconnection between them; for instance, from formula (18) it is easy to derive the argument principle (Sec. II.3.7) of the theory of analytic func-

tions (let the reader do it!), which elucidates the geometric meaning of the principle.

What has been said is directly related to the theory of autonomous systems since every such system is determined by a vector field of velocities. Even if the field is continuous it may have singularities at the rest points of the system. It is evident that the index of every cycle (L) of the system relative to the field of velocities is always equal to unity (why?). Consequently, by Property 2, *there is at least one rest point inside every cycle* (L). More precisely, formula (18) shows that the sum of the indices of the rest points lying within (L) equals unity. Therefore, if these points are only of the types described in Sec. 3, the number of the saddle points must be less by unity than the total number of the rest points of other types.

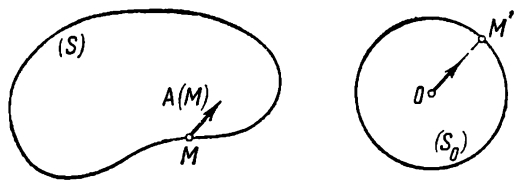


Fig. 145

In conclusion we note that there are singular points of a more complicated structure (see Sec. 4) for which the index can be equal to any integer (let the reader consider this question).

The notion of the index of a singular point of a vector field can be defined for a Euclidean space of arbitrary dimensions. For definiteness, let us speak about the three-dimensional geometric space. Consider a vector field $\mathbf{A}(x, y, z)$ and choose a closed surface (S) not passing through singular points of the field. With every point $M \in (S)$ we associate a point $M' \in (S_0)$, where (S_0) is a fixed sphere of unit radius and centre O, according to the rule $\overrightarrow{OM'} = \frac{\mathbf{A}(M)}{|\mathbf{A}(M)|}$ (see Fig. 145). The index $\gamma[\mathbf{A}, (S)]$ of the surface (S) relative to the field $\mathbf{A}$ is defined as the number of times the sphere S_0 is covered by the points M' under this mapping of (S) into (S_0) as M runs through (S) once, on condition that every part of (S_0) covered under the mapping is counted with the sign plus or minus depending on whether the mapping preserves the orientation of that part or not. Here the surfaces (S) and (S_0) are supposed to be oriented in the natural manner, that is in the sense that their outer sides face infinity. In the case of an arbitrary dimension n this means that each of these $(n-1)$-dimensional surfaces has the orientation coherent with the natural

orientation of the n-dimensional figure bounded by it (e.g. [107], Secs. XVI.20 and 29), the natural orientation of E_n being specified by the ordered sequence of points $(1, 0, \ldots, 0)$, $(0, 1, \ldots, 0)$, $\ldots$, $(0, 0, \ldots, 1)$, $(0, 0, \ldots, 0)$. (Let the reader analyse this construction for $n=2$ and $n=3$.)

The index thus introduced possesses all the properties enumerated above for plane fields. It is also possible to derive a formula analogous to (17): if

$$\mathbf{A}=\sum_{j=1}^{n} A_j(x_1, x_2, \ldots, x_n)\,\mathbf{e}_j$$

where $\mathbf{e}_j$ $(j=1, 2, \ldots, n)$ is the unit vector of the x_j-axis, then

$$\gamma[\mathbf{A}, (S)]=\frac{1}{\omega_{n-1}}\int_{(S)}\sum_{j=1}^{n}\begin{vmatrix} \frac{\partial a_1}{\partial x_1} & \cdots & \frac{\partial a_1}{\partial x_{j-1}} & a_1 & \frac{\partial a_1}{\partial x_{j+1}} & \cdots & \frac{\partial a_1}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial a_n}{\partial x_1} & \cdots & \frac{\partial a_n}{\partial x_{j-1}} & a_n & \frac{\partial a_n}{\partial x_{j+1}} & \cdots & \frac{\partial a_n}{\partial x_n} \end{vmatrix}\frac{\cos(\mathbf{n}, \widehat{x_j})}{|\mathbf{A}|^n}\,dS$$

where $\mathbf{n}$ is the unit vector of the outer normal to (S) and ω_{n-1} the $(n-1)$-dimensional measure of (S_0) which can be computed by the recurrence formula

$$\omega_{n-1}=2\frac{n}{n-1}\int_0^{\frac{\pi}{2}}\sin^n\varphi\,d\varphi\,\omega_{n-2}\quad (n=2, 3, \ldots;\ \omega_0=2),$$

$$\int_0^{\frac{\pi}{2}}\sin^n\varphi\,d\varphi=\begin{cases}\frac{n-1}{n}\cdot\frac{n-3}{n-2}\cdots\frac{1}{2}\cdot\frac{\pi}{2} & \text{for even } n\\ \frac{n-1}{n}\cdot\frac{n-3}{n-2}\cdots\frac{2}{3} & \text{for odd } n\end{cases}$$

In the general case the notion of the index of a singular point is also introduced by analogy with the case of a plane field.

7. Rest Points in Space. The behaviour of trajectories in a small neighbourhood of a rest point in a space of any dimension n is easily investigated in structurally stable cases. We shall confine ourselves to the three-dimensional $x_1x_2x_3$-space, the course of the investigation being analogous for spaces of higher dimensions.

As in Sec. 3, we assume that the rest point in question is at the origin and pass to the linearized system to obtain a new system of type (10) in which

$$\mathbf{r}=\begin{pmatrix}x_1\\ x_2\\ x_3\end{pmatrix},\qquad \mathbf{A}=\left(\frac{\partial f_i}{\partial x_j}\right)_{x=0}$$

A transformation $\mathbf{r} = \mathbf{Tr}'$ leads to system of form (11); this again is connected with the problem of reducing the matrix $\mathbf{A}$ to a simpler form by means of a real transformation matrix.

The ultimate result depends on the roots λ_1, λ_2, λ_3 of the characteristic equation

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0 \tag{19}$$

If all the roots are real and distinct the matrix can be transformed to the diagonal form; system (11) then is written as $\dot{x}_j = \lambda_j x_j$, its general solution being

$$x_j = Ce_j^{\lambda_j t} \qquad (j = 1, 2, 3) \tag{20}$$

For instance, if $\lambda_1 < \lambda_2 < \lambda_3 < 0$, it is easy to see that for $t \longrightarrow \infty$ every integral curve enters the rest point, the direction along which it goes coinciding with the x_3-axis for $C_3 \neq 0$, with the x_2-axis for $C_3 = 0$, $C_2 \neq 0$ and with the x_1-axis for $C_3 = C_2 = 0$, $C_1 \neq 0$. Thus, we have a stable nodal point with three distinct directions along which the trajectories enter it. If $\lambda_j > 0$, $j = 1, 2, 3$, the nodal point is unstable. In the case of multiple roots with the same sign we also have a node but the number of directions of entrance may be reduced. For $n > 3$ there is also a nodal point if all the roots are of the same sign.

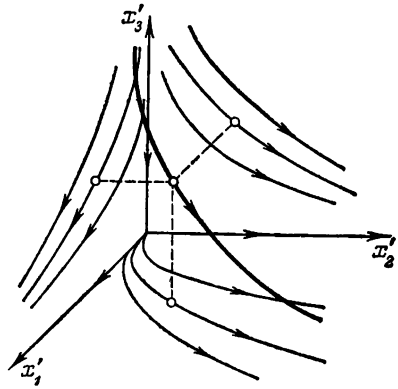


Fig. 146

If the roots are real and distinct but not of one sign formula (20) is again valid but the geometric meaning changes; for instance, in Fig. 146 we see the projections of the trajectories on the coordinate planes shown for the case $\lambda_3 < 0 < \lambda_1 < \lambda_2$ (these projections are themselves trajectories) and an essentially spatial (twisted) trajectory depicted in heavy line. In this case the origin is a saddle point (which also appears for multiple real roots with different signs). Only the trajectories in the $x_1'x_2'$-plane (which is a separatrix; see Sec. 2) and those going along the x_3'-axis enter the rest point (the former for $t \longrightarrow -\infty$ and the latter for $t \longrightarrow \infty$), all the other trajectories not passing through it. A saddle point is always unstable.

If among the roots of equation (19) there is a pair of conjugate complex roots it is possible, by Sec. IV.3.4, to reduce system (11) to the form

$$\dot{x}_1' = \mu x_1' + \nu x_2', \qquad \dot{x}_2' = -\nu x_1' + \mu x_2', \qquad \dot{x}_3' = \lambda_3 x_3'$$

Introducing polar coordinates in the $x_1'x_2'$-plane and integrating the resultant equations we obtain

$$\rho = C_1 e^{\mu t}, \quad \varphi = -\nu t + C_2, \quad x_3' = C_3 e^{\lambda_3 t}$$

If μ and λ_3 are of the same sign the location of the trajectories resembles both a nodal and a focal point; a **nodal-focal point** of

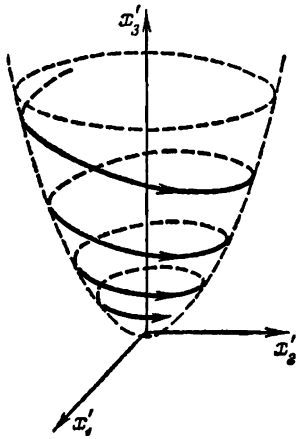


Fig. 147

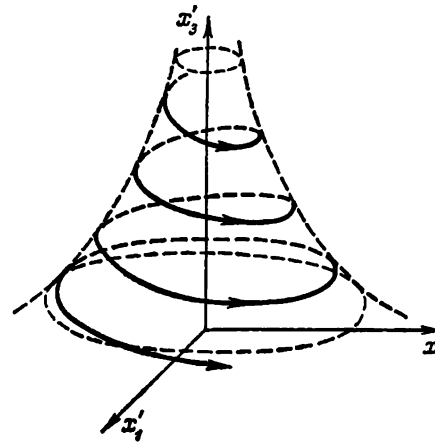


Fig. 148

this kind can be stable (Fig. 147) or unstable. If μ and λ_3 are of opposite signs there is a **saddle-focal point** at the origin (Fig. 148) which is always unstable. For $\mu = 0$ we obtain a picture as in Fig. 149.

In all cases except the one corresponding to Fig. 149 the addition of higher-order terms does not change the general structure of the trajectories while a point of the type shown in the figure may turn into a node-focus or saddle-focus. Besides, there exists a wide variety of different cases when the matrix **A** is degenerate and, in particular, equal to zero.

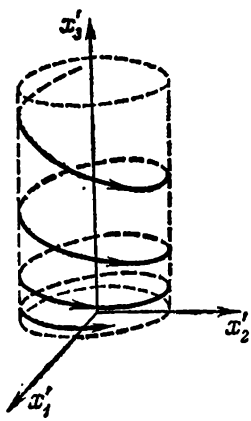


Fig. 149

For $n > 3$ the possible cases are similar, the only new feature being that when the matrix **A** possesses several pairs of imaginary eigenvalues there can be a superposition of several rotations, and for pure imaginary roots there may appear quasi-periodic solutions (Sec. 1.7).

If $\det \mathbf{A} \neq 0$ the transformation $\mathbf{r} \longrightarrow \mathbf{Ar}$ defines an affine mapping of the space into itself under which the orientation is preser-

ved if $\det \mathbf{A} > 0$ or is changed to the opposite if $\det \mathbf{A} < 0$. Hence, the definition of the index of a singular point (see Sec. 6) implies that if the condition $\det\left(\frac{\partial f_i}{\partial x_j}\right) > 0$ $\left(\det\left(\frac{\partial f_i}{\partial x_j}\right) < 0\right)$ holds at the rest point its index is equal to 1 (−1). If $\det\left(\frac{\partial f_i}{\partial x_j}\right) = 0$ the index may assume other values as well.

For the case of dimension $n > 2$ there is an analogue of the theorem on the existence of a rest point in the plane inside a closed trajectory (see Sec. 6). Consider, in an n-dimensional space where an autonomous system of type (2) is defined, a spatial figure (T) topologically equivalent to an n-dimensional ball. The bounding surface (S) of (T) is assumed to be smooth or piecewise-smooth (more precisely, (S) may have singular manifolds of dimension not higher than $n-2$). Suppose that the vectors of the corresponding field of velocities are at each point of (S) distinct from zero and directed toward the interior of (T) or at least along a tangent to (S); this means that for $x_0 \in (T)$ and $t \geqslant 0$ the condition $x(t, x_0) \in (T)$ holds. Then there is at least one rest point inside (T). Virtually, it follows from the hypotheses that the index of the surface (S) relative to the field of velocities is equal to unity, whence the assertion follows (why?).

If for $n = 3$ we replace the ball by a torus the above assertion is no longer true because even when all the other hypotheses hold a domain of that kind may contain no rest point. At the same time it is allowable to take as (T) a figure topologically equivalent to a ball with two or more "handles" (a torus is topologically equivalent to a ball with one handle). In courses on *algebraic topology* it is proved that for a figure (T) of an arbitrary dimension to contain at least one rest point of any autonomous system defined in (T) whose trajectories cannot leave (T) as time t increases it is sufficient that the so-called **Euler characteristic** of (T) be nonzero. Here we shall not go into particulars and only mention that this characteristic assumes integral values and is a topological invariant of (T). The Euler characteristic of an n-dimensional ball is equal to 1 while for an $(n-1)$-dimensional sphere it is equal to $1-(-1)^n$. For a three-dimensional ball with k handles it is equal to $1-k$ and for the surface of this figure to $2-2k$.

8. Cycles in Space. Every cycle of system (2) is determined by a periodic solution $x_0(t)$ of the system. Let T be the period of such a solution. To study the behaviour of the solutions in a small neighbourhood of the cycle we make the substitution $\mathbf{x} = \mathbf{x}_0(t) + \xi$, expand the right-hand members in powers of the projections of ξ and discard, for small $|\xi|$, the terms whose order of smallness relative to $|\xi|$ is higher than the first. This results in the linea-

rized system

$$\dot{\xi} = f'_x(x_0(t))\,\xi \tag{21}$$

with a periodic coefficient matrix of period T. (We remind the reader that, by the definition, f'_x is understood as the matrix $\left(\frac{\partial f_j}{\partial x_k}\right)$.) According to Sec. 1.2, the behaviour of the solutions of system (21) for $t \to \infty$ is specified by its characteristic multipliers ρ_j which can be computed in concrete problems by constructing the monodromy matrix (Sec. 1.2) with the aid of numerical integration and determining its eigenvalues. As in the study of rest points, in basic cases the behaviour of the solutions of the original system for small $|\xi|$ is the same as in the linearized system (21). But it should be taken into account that an autonomous system of type (2) having a periodic solution $x_0(t)$ possesses the one-parameter family $x_0(t+\lambda) = x_0(t) + \lambda \dot{x}_0(t) + \ldots$ of periodic solutions, and therefore the corresponding system of type (21) has $\dot{x}_0(t)$ as one of its particular solutions to which the multiplier $\rho = 1$ corresponds. Therefore, the classification of the cycles is based on the properties of the remaining $n-1$ multipliers, and the resulting conclusions are formulated as follows.

If all these $n-1$ multipliers of system (21) are less than unity in their moduli the corresponding cycle attracts the trajectories in the sense that all the trajectories lying sufficiently close to it tend to this cycle for $t \to \infty$ approaching it spirally (that is winding about it). Moreover, as was proved by A. M. Lyapunov, for every such trajectory $x(t)$ there exists τ (time shift) such that the distance between the points $x(t+\tau)$ and $x_0(t)$ tends to zero as $t \to \infty$, the constant τ being, in the general case, different for different trajectories. If the moduli of all these multipliers exceed unity the cycle $x_0(t)$ repulses the trajectories. Finally, if there are multipliers whose moduli both exceed unity and are less than unity, the cycle $x_0(t)$ is unstable and resembles, in its character, a saddle point because there is a manifold of dimension less than n filled with trajectories which tend to the cycle as $t \to \infty$. The cases when the system possesses a number of multipliers whose moduli are equal to unity while the moduli of the other multipliers are all greater or less than unity are rather complicated (except the trivial case mentioned at the beginning of the foregoing paragraph), and we shall not dwell on them here.

The above properties imply a simple test for stability applicable to a cycle of an autonomous system of form (5) in the plane. Let the equations of the cycle be $x = \varphi(t)$, $y = \psi(t)$ where the functions $\varphi(t)$ and $\psi(t)$ are of period T. The corresponding linearized

system is written as

$$\dot{\xi} = P'_x(\varphi(t),\ \psi(t))\,\xi + P'_y(\varphi(t),\ \psi(t))\,\eta,$$
$$\dot{\eta} = Q'_x(\varphi(t),\ \psi(t))\,\xi + Q'_y(\varphi(t),\ \psi(t))\,\eta$$

Let us use formula (1.18). According to it, since one of the multipliers is equal to unity, the limit cycle in question is stable if

$$\int_0^T [P'_x(\varphi(t),\ \psi(t)) + Q'_y(\varphi(t),\ \psi(t))]\,dt < 0 \qquad (22)$$

If this integral is positive the cycle is unstable. This criterion was established by H. Poincaré.

The method of point mappings (Sec. 5) can also be applied to cycles lying in a many-dimensional space. We shall demonstrate this application for the three-dimensional space. Suppose that we have a surface (S), described by an equation $F(x) = 0$, with no points of tangency to the trajectories of the system under consideration (Fig. 150), on which a coordinate system α_1, α_2 is chosen (we shall write $\alpha = (\alpha_1, \alpha_2)$), that is $x = \varphi(\alpha)$ $(x = (x_1, x_2, x_3))$. If the trajectory starting at every point $\varphi(\alpha) \in (S)$ intersects the surface (S) repeatedly at a point $x' = \varphi(\alpha') = x(t', \varphi(\alpha))$ (which can be found from the equation $F(x(t, \varphi(\alpha))) = 0$) we have a Poincaré's mapping $\alpha'(\alpha)$ of (S) into itself. A cycle (provided it exists) is specified by a value $\alpha = \alpha_0$ satisfying the equation $\alpha'_0 = \alpha_0$, i.e. $\alpha'(\alpha_0) = \alpha_0$ (see Fig. 150), the properties of the cycle depending on the character of the mapping $\alpha'(\alpha)$ in the vicinity of α_0. In particular, it can be shown that the eigenvalues of the matrix $\left(\frac{\partial \alpha'}{\partial \alpha}\right)_{\alpha_0}$ coincide with the multipliers of system (21) except the trivial multiplier $\rho = 1$. These considerations again imply the conclusions drawn in the preceding paragraph.

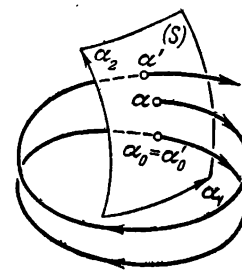


Fig. 150

The method of mappings can be applied practically when there is some information about the location of the cycle we are interested in. The value $\alpha' = \varphi(\alpha)$ can be determined, for every concrete α, with the aid of numerical integration of system (2) and the determination of the sign of $F(x)$ along the trajectory. Hence, attracting cycles can be found by means of the simple iteration scheme $\alpha_{n+1} = \varphi(\alpha_n)$ and repulsing ones by the same procedure with the integration for decreasing t. Cycles of other types can be found by solving the equation $\alpha = \varphi(\alpha)$ numerically with the help of Newton's method (e.g. see [107], Sec. XII.12). The values

of the derivatives $\frac{\partial \varphi_j}{\partial \alpha_k}$ needed for this method can be computed by using variation equations (e.g. see [107], Sec. XV.28) or by applying formulas for numerical differentiation.

When constructing a cycle one must take into account that a nonclosed trajectory in an n-dimensional space bounded for $t \to \infty$ and having no rest points in its ω-limit set may not be convergent spirally to a cycle when $n \geqslant 3$, in contrast to the case $n=2$, and its asymptotic behaviour may be more complicated.

There is an approach developed by H. Poincaré applicable to constructing and investigating cycles. Suppose we are given a system of form (2) with one or more parameters:

$$\dot{\mathbf{x}} = \mathbf{f}(x, \mu) \qquad (\mu = (\mu_1, \ldots, \mu_m)) \tag{23}$$

The system being autonomous, its general solution is expressible as

$$x = F(t - t_0, C, \mu), \qquad (C = (C_1, C_2, \ldots, C_{n-1}))$$

Let, for a certain $\mu = \mu_0$, a periodic solution $F(t - t_0, C_0, \mu_0)$, with period T_0, of system (23) be known. Then it is natural to pose the question as to whether there are periodic solutions for the values of μ close to μ_0.

Denoting by T the period (yet unknown!) of such a solution and taking into account that the system is autonomous we can write the periodicity condition for the solution in the form

$$F(T, C, \mu) = F(0, C, \mu), \quad \text{that is} \quad F(T, C, \mu) - F(0, C, \mu) = 0 \tag{24}$$

By the hypothesis, for $T = T_0$, $C = C_0$ and $\mu = \mu_0$ the latter system of equations is satisfied. Consequently, according to the theorem on existence of an implicit function (e.g. see [107], Sec. IX.13), system (24) is uniquely resolvable with respect to T and C for small $|\mu - \mu_0|$ if, for these values, the condition

$$\frac{\partial (F(T, C, \mu) - F(0, C, \mu))}{\partial (T, C)} \neq 0 \tag{25}$$

holds (the expression on the left-hand side is a Jacobian of order n). Then system (23) possesses exactly one cycle close to the given one.

Condition (25) being considered for the given cycle, it is allowable to take as F, in (25), the solution of linearized system (23) for $\mu - \mu_0$. Although the solution of the linearized system can also be found explicitly only in some special cases, the linearization may prove useful for testing condition (25) numerically or, which is more important, for numerical construction of the functions $T(\mu)$ and $C(\mu)$ for the given cycle, when $|\mu - \mu_0|$ is small, and exten-

sion of the functions thus obtained according to the general method given in Sec. IV.4.11.

Here we shall not deal with the case when condition (25) is violated to which extensive research has been devoted. This case may arise when for $\mu = \mu_0$ the system has a multiple cycle which splits into several cycles or disappears as μ is varied or when the system has a manifold of cycles. For these problems we refer the reader to [91].

9. Structurally Stable Systems. If an object possesses the property of preserving its basic characteristics when the parameters it depends on are given sufficiently small variations, we say that the object is *structurally stable*. Such objects (systems) are important in practical aspects because in reality only approximate values of the parameters can be guaranteed and therefore structurally unstable systems cannot be realized unless special measures are taken for fixing these values.

This of course does not mean that structurally unstable cases do not deserve investigation. If, for instance, the formulation of a problem involves a continuously varying parameter λ, the condition $\lambda = 0$ (or $\lambda = \lambda_0$ where λ_0 is any fixed value) is not structurally stable but it may be essential for the schematization of the problem (for example, if λ is a coefficient of friction this condition means that we consider an idealized process without dissipation of energy, and the like). Furthermore, the investigation of the case $\lambda = 0$ may lead to an extension of the results thus obtained to the structurally stable case $\lambda \neq 0$ with the aid of the small-parameter method or some other continuation scheme. If there is a family of problems under consideration in which a parameter λ varies continuously passing from negative values to positive ones it may turn out to be necessary to include the case $\lambda = 0$ as well. This approach is also applicable if λ is known to be small with probability close to 1, and so on.

The concrete definition of structural stability varies depending on the class of systems in question, on the properties regarded as fundamental which should remain invariant, on the range of the admissible variations of the parameters, and so on. The Soviet scientists A. A. Andronov (1901-1952), a specialist in the theory of vibrations, and L. S. Pontryagin (born in 1908), a mathematician, introduced and investigated the notion of structural stability for autonomous differential systems in the plane (extended, as in the theory of analytic functions, by the addition of the point at infinity; see Sec. II.1.7). According to their definition, a system of this type is said to be structurally stable if for any sufficiently small variations of the right members of the differential equations and their derivatives the general arrangement of the trajectories is invariant in the large or, more precisely, if there is a differentiable

in both directions one-to-one mapping of the plane onto itself under which all the trajectories of the perturbed system go into the trajectories of the original system. It was found that for structural stability of this kind it is necessary and sufficient that the following conditions hold:

(1) there is a finite number of rest points each of which, in linear approximation, belongs to the first three types described in Sec. 3;

(2) there is a finite number of cycles each of which possesses a multiplier distinct from unity;

(3) there are no separatrices going from one saddle point to another.

(Let the reader carefully examine these conditions.)

Every sufficiently small perturbation of a structurally stable system results in a new structurally stable system while, generally speaking, a system without structural stability can be made structurally stable by means of a small variation. (This can be compared to the well-known situation in analysis: every function which is not identically constant retains this property when it is given a sufficiently small variation whereas a constant function can be changed into a nonconstant one by means of an arbitrarily small variation.)

Every stable point and every stable cycle has a *domain of attraction* in the phase plane, covered with trajectories having this rest point or cycle as its ω-limit set. The phase plane of a structurally stable autonomous system is broken up into such domains by separatrices entering saddle points.

For autonomous systems in an n-dimensional space ($n \geqslant 3$) it is possible to formulate some necessary conditions for structural stability, similar to the above Conditions 1-3, which are not sufficient. The necessary and sufficient conditions for structural stability are at present unknown for the general case. In conclusion we mention that recently it has been shown that there exist structurally unstable systems that cannot be made structurally stable by means of an arbitrarily small variation.

10. Discontinuous Systems. In applied problems we often deal with cases when a system of form (2) has continuous right members defined on both sides of an $(n-1)$-dimensional surface (S) while on (S) they have discontinuities. This is connected with some new features which we shall discuss for the simpler case $n=3$.

Let us regard the situation as if we are given two autonomous systems (with continuous right-hand sides) defined in the vicinity of (S) so that on one side of (S) we must use one of the systems and on the opposite side the other system. The simplest case occurs when the trajectries of both systems go in the same direction from (S) at each point belonging to (S) (see Fig. 151). Then,

generally speaking, the trajectories of the discontinuous system in question (regarded as an aggregate of the two given systems) have corner points on (S), which does not lead to any further complications. In this case we construct the trajectories of both systems in such a way that the terminal values achieved on (S) by one of them serve as initial values for the other system.

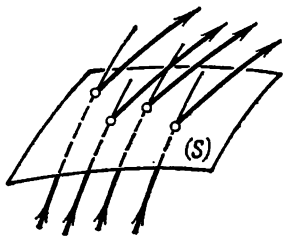


Fig. 151

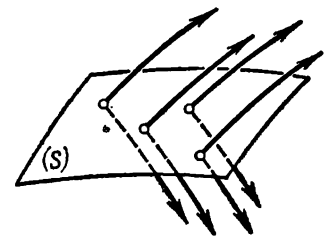


Fig. 152

The case when the trajectories of both systems are directed away from (S) at the points belonging to (S) (see Fig. 152) encounters some difficulties when the initial point is taken on (S) since a trajectory of this type can be continued in either direction from (S), and thus to every initial point on (S) there correspond two different trajectories. But the trajectories extended from (S) (or those whose initial points do not lie on (S)) cannot intersect (S) repeatedly (why?), and therefore there are no other complications.

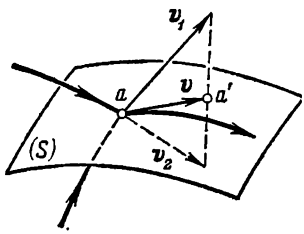


Fig. 153

The most interesting case appears when in the vicinity of (S) the trajectories of both systems are directed toward (S) (Fig. 153). In this case, after a trajectory reaches (S), it cannot leave it in either direction. On the other hand, the limiting velocity vectors on (S) form nonzero angles with (S), and therefore the trajectory seems to "disappear" after it has reached (S).

To continue the trajectories in a meaningful way we use the following notion of the generalized solution of system (2). Let us lay off, from each point $a \in (S)$, the limiting velocity vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ corresponding to both sides of (S) (see Fig. 153), connect their tips with a line segment and choose a point a' on it such that the vector $\mathbf{v} = \overrightarrow{aa'}$ is tangent to (S). This construction specifies a field of tangent vectors $\mathbf{v}$ on (S), and every trajectory (vector line) of this field can naturally be regarded as an extension of

the corresponding trajectory which has reached (S) (we thus obtain the so-called "**sliding regime**"). A trajectory continued in this way can either remain in the sliding regime on (S) for all the time or leave the surface (S) (which may be caused by a perturbation or some other factor).

If the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ (see Fig. 153) are opposite to each other we have $\mathbf{v}=\mathbf{0}$, and, consequently, in this case the point a is a stationary point for the sliding regime. There can also exist

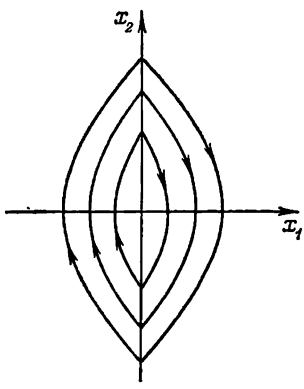


Fig. 154

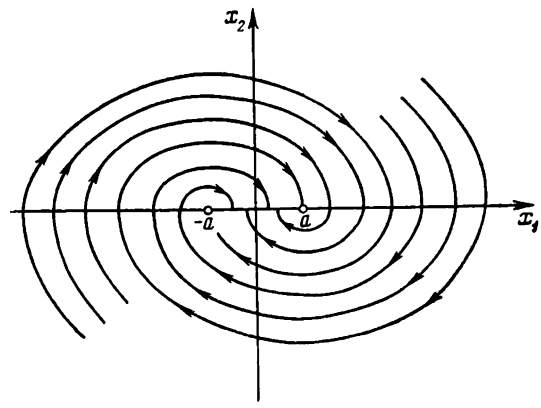


Fig. 155

a region on (S) consisting entirely of such points; such a region is spoken of as an "**inertia zone**".

Let us consider some examples involving the simplest discontinuous function

$$\operatorname{sgn} x=\begin{cases}-1 \text{ for } -\infty<x<0\\ 0 \text{ for } x=0\\ 1 \text{ for } 0<x<\infty\end{cases}$$

known as the **signum function** which is connected with the **Heaviside unit function (step function)** $e(x)$ (e.g. see [107], Sec. XIV.25) by the formula $2e(x)=1+\operatorname{sgn} x$. Such functions usually appear in "switching systems" whose differential law of variation changes in jump-like manner when a certain state of the system has been reached.

Consider the equation $\ddot{x}+a\operatorname{sgn} x=0$ equivalent to the system

$$\dot{x}_1=x_2,\quad \dot{x}_2=-a\operatorname{sgn} x_1 \qquad (x_1\equiv x)$$

The direct integration for $x>0$ and $x<0$ results, for $a>0$, in the **phase diagram** shown in Fig. 154 (check it up!). On the line of discontinuity $x_1=0$ the trajectories behave as in Fig. 151. Let the reader consider the case $a<0$.

The system

$$\dot{x}_1 = x_2,\ \dot{x}_2 = -\omega^2 x_1 - a \operatorname{sgn} x_2$$

has, for $a > 0$, the phase diagram shown in Fig. 155 (let the reader check the construction). The interval $-a \leqslant x_1 \leqslant a$ of the x_1-axis is an inertia zone at whose every point the system can be in a state of rest. For any given initial conditions the trajectory of the system, after a finite number of oscillations, arrives, in a finite time, at this zone where it becomes stationary. It should be noted that if, after the system reaches such a stationary state, it is given a random perturbation, it has a tendency to pass to the state specified by the values $x_1 = x_2 = 0$ (why?).

Now consider an example in which the right members of the system are themselves continuous but their derivatives are discontinuous. Take an

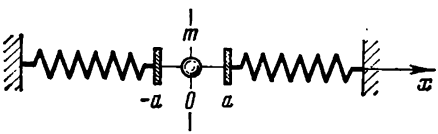


Fig. 156

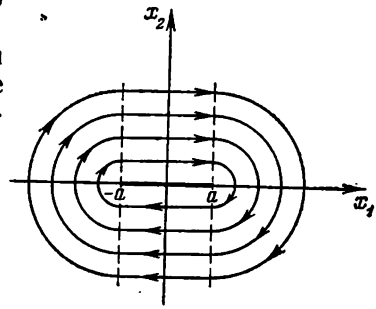


Fig. 157

oscillator with clearance whose simplest scheme is shown in Fig. 156. The equation of motion of such an oscillator is written as $m\ddot{x} + \varphi(x) = 0$ where

$$\varphi(x) = \begin{cases} k(x+a) & \text{for } -\infty < x < -a \\ 0 & \text{for } -a < x < a \\ k(x-a) & \text{for } a < x < \infty \end{cases}$$

Accordingly, the corresponding system of equations has the form

$$\dot{x}_1 = \frac{1}{m} x_2,\ \dot{x}_2 = -\varphi(x_1)$$

its phase diagram being shown in Fig. 157. The interval $-a \leqslant x_1 \leqslant a$ of the x_1-axis consists of rest points and thus is an inertia zone; but this inertia zone differs, in its nature, from the one in Fig. 155: in this case arbitrarily small perturbations lead to slow oscillations with finite amplitude. (What is the physical significance of these oscillations?)

Many similar and more complicated examples of discontinuous systems are considered in [4].

In conclusion we remark that a switching system may involve several autonomous systems of type (2) defined in a given domain

and alternating according to a certain law. A trajectory of the point representing a state of such a switching system may have points of self-intersection.

11. Systems on Manifolds. Up till now we assumed that system (2) is defined in the whole space E_n or in a domain of E_n. But such a system may also be defined on an arbitrary n-dimensional manifold with generalized coordinates $x_1, x_2, \ldots, x_n$.

For instance, suppose that we are given a system with one degree of freedom, its coordinate φ being interpreted as an angle whose values φ and $\varphi+2\pi$ determine the same state of the system.

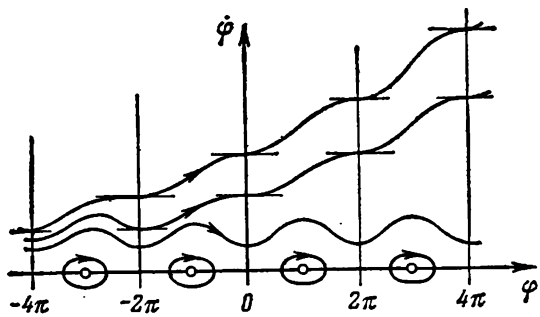


Fig. 158

Then, to describe the configuration of the system, it suffices to take the interval $0 \leqslant \varphi \leqslant 2\pi$ to whose end points there corresponds the same state. Therefore, the configuration manifold of the system is a circle (more precisely, it is topologically equivalent to a circle; let the reader analyse this property).

If this system is described by an equation $\ddot{\varphi} = f(\varphi, \dot{\varphi})$ the function f must be periodic in φ with period 2π. The $\varphi\,\dot{\varphi}$-plane serves as the phase space of the system to whose every point there corresponds a certain state of the system, i.e. its coordinate and velocity. As has been said, the configuration set of the system is a circle and, consequently, since the set of velocities is, in this case, an infinite straight line, the phase space can be interpreted as a circular cylinder which is the *topological product* of a circle and a straight line. Hence, in the case under consideration the vector field of velocities, the trajectories, rest points, etc. can be represented on the cylinder. This system can also be represented on an ordinary plane with coordinates φ and $\dot{\varphi}$ but one must bear in mind that in this interpretation all the strips of width 2π parallel to the $\dot{\varphi}$-axis (see Fig. 158) are considered to be identified as if the plane is wound around a circular cylinder for which the circumference of the normal section is equal to 2π (for instance,

let the reader check that Fig. 158 shows one rest point, two closed and one nonclosed trajectories).

Similarly, if all the right-hand members of general system (1) are periodic in every variable x_j with period $a_j (j = 1, 2, \ldots, n)$ its phase manifold is topologically equivalent to the product of n circles. In particular, for $n = 2$ we have the surface of a torus since its every point can be characterized in a one-to-one manner by certain values of two angular coordinates (how?). In this case we simply speak of an autonomous system defined on a torus (accordingly, for an arbitrary n we speak of a system on an n-dimensional torus) meaning that its phase manifold is topologically equivalent to the surface of a torus. Such a system can also be represented on an ordinary x_1x_2-plane divided into rectangles of dimensions a_1 and a_2, all the rectangles being identified with each other.

An autonomous system can be defined on a sphere of arbitrary dimension and the like. There can also be more complicated phase manifolds; for instance, in the problem of the motion of a rigid body about a fixed point the state space is the three-dimensional manifold of all orthogonal matrices of the third order with determinant equal to unity (why?). The structure of such manifolds is studied in courses of *combinatorial topology* (which in modern mathematics is regarded as a part of *algebraic topology*, both terms being sometimes used as synonyms).

Tori and spheres are examples of compact manifolds while a cylinder is noncompact. The α- and ω-limit sets of every trajectory in a compact manifold always contain at least one point (why?), are closed, composed of entire trajectories and connected (Sec. 2).

The theory of autonomous systems on general manifolds involves some new features in comparison with the theory of systems defined in the space E_n. To elucidate what has been said let us dwell on the case $n = 2$. First of all, a closed line on an arbitrary two-dimensional manifold does not necessarily divide it into a finite and an infinite part; it may also divide it into two finite portions (as every closed trajectory on a sphere does) or into two infinite parts (see the upper cycle in Fig. 158) or may not divide it into parts at all (this is the case for a meridian or a parallel on a torus). That is why the existence of a cycle on a cylinder or torus does not imply, in the general case, the existence of a rest point. On the other hand, the results presented in Sec. 7 show that an autonomous system defined on a sphere always possesses at least one rest point irrespective of the existence of cycles, but this does not of course apply to the torus. Hence, although the local properties of systems on two-dimensional manifolds (such as the behaviour o the corresponding vector field in the vicinity of a rest point, etc.)

are the same as in the case of a plane, their structure in the large (global properties) is substantially dependent on the topological nature of the manifold.

12. Systems with an Integral Invariant. For definiteness, we shall consider a system of form (2) defined in E_n although the discussion below also applies to systems on arbitrary manifolds.

Let (V) be a domain in E_n. We shall designate by the symbol $(V)_t$ the region formed of all the points at which the points of (V) moving along the trajectories of the system arrive after a time t. In other words, $(V)_t$ is the collection of all the points $x(t, x_0)$ $(x_0 \in (V))$, where $x = x(t, x_0)$ is the solution of system (2) satisfying the initial condition $x|_{t=0} = x_0$. If there exists a function $\rho(x)$ defined in E_n such that the integral

$$\int_{(V)_t} \rho(x)\, dx \qquad (26)$$

is time-independent for any finite domain (V) we call integral (26) an **integral invariant** of system (2), the function $\rho(x)$ being referred to as the **density of the integral invariant.** The geometric meaning of this definition is demonstrated in Fig. 159: the volume of the cylindrical body whose base is the moving region $(V)_t$ remains invariant.

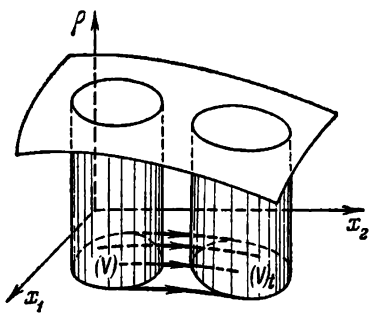


Fig. 159

To derive an equation for a density of an integral invariant we note that for integral (26) to be invariant it is necessary and sufficient that its every element be invariant, that is

$$D(\rho(x)\, dV) = 0 \qquad (27)$$

where D denotes the operation of taking the differential resulting from the shift along the trajectory in a time dt. According to the meaning of the divergence of a field of velocities (e.g. see [107], Sec. XVI.23), we can write $D(\rho dV) = D\rho\, dV + \rho D(dV) = \operatorname{grad} \rho \cdot \dot{\mathbf{x}}\, dt\, dV + \rho \operatorname{div} \dot{\mathbf{x}}\, dV\, dt$. Therefore, by (2) and (27), we obtain the desired equation

$$\operatorname{div}(\rho \mathbf{f}(x)) = \operatorname{grad} \rho \cdot \mathbf{f}(x) + \rho \operatorname{div} \mathbf{f}(x) = 0 \qquad (28)$$

expressing a necessary and sufficient condition for a given function $\rho(x)$ to be a density of an integral invariant of system (2).

If $\rho \equiv \text{const}$ is such a density the n-dimensional volume of every region is preserved under the shift along the trajectories. It follows from (28) that this is the case if and only if $\operatorname{div} \mathbf{f}(x) = 0$. In par-

ticular, this property is characteristic of canonical systems

$$\dot{x}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial x_i}$$

$$(i = 1, 2, \ldots, n; \; H = H(x_1, \ldots, x_n, p_1, \ldots, p_n))$$

(see Sec. VI.3.1) regarded in the $2n$-dimensional phase space of the variables $x = (x_1, \ldots, x_n)$, $p = (p_1, \ldots, p_n)$ (check up this assertion!). Hence, the $2n$-dimensional volume in this space is preserved under the shift along the trajectories (this proposition is known as **Liouville's theorem**).

Suppose that a function $\rho(x)$ is a density of an integral invariant of system (2). Let us introduce the measure

$$\mu_{(V)} = \int_{(V)} \rho \, dV$$

in the space $x_1, \ldots, x_n$ (for the general notion of measure see e.g. [107], Sec. XVI.19). According to the above, this is an **invariant measure** with respect to the shift along the trajectories. The most interesting case is when $\rho(x) > 0$.

Here and henceforth we shall assume that the density $\rho(x)$ is continuous and satisfies the condition $0 < \rho(x) < \infty$. It is easy to show that any system of type (2) possesses an integral invariant in the vicinity of its every point which is not a stationary point. But this result is not of great significance since we are particularly interested in systems having an integral invariant defined in the whole phase space.

It should be stressed that by far not every autonomous system possesses an integral invariant; in real problems the existence of such an invariant is in a certain sense equivalent to the property of conservation of energy.

Now suppose that system (2) with integral invariant has a stationary point x_0. Then, putting $x = x_0$ in (28), we see that $\operatorname{div} \mathbf{f}(x_0) = 0$. The left member of this relation is equal to the sum of the roots of the characteristic equation of the linearized system (2) at the point x_0 (why?) and, consequently, if there are roots with positive real parts there must also be roots with negative real parts and vice versa. Therefore, x_0 cannot be a structurally stable rest point attracting or repulsing trajectories; for instance, among the types of rest points enumerated in Sec. 3, only saddle points and centres can be at present in a system with integral invariant. Stable nodal and focal points are characteristic of systems with dissipation of energy while unstable ones of systems with an external source of energy.

The same conclusions can be drawn on the basis of the following visual considerations. For instance, take a stable nodal point

(see Fig. 160) and a region (V) in its vicinity. For $t \to \infty$ this region degenerates into a point, which indicates that integral (26) cannot be invariant (why?). Analogous considerations show that a system with integral invariant cannot have limit cycles attracting or repulsing trajectories and, generally, it cannot have attracting or repulsing manifolds of dimension less than that of the phase space.

13. Ergodicity. Let an autonomous system with an integral invariant be defined on a compact manifold (M) (such systems are sometimes referred to as **dynamic systems**). As was said, a system of this kind determines a flow on (M), that is a transformation $x \to x_t$ (Sec. 1), and an invariant measure μ with respect to the flow.

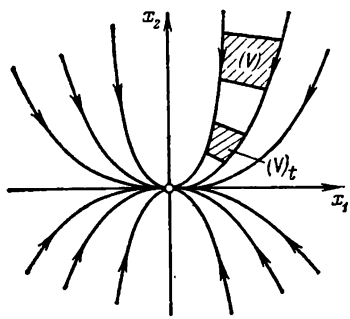


Fig. 160

Now consider an arbitrary finite function $f(x)$ defined on (M), and take its mean (average) value $\bar{f}(x)$ along the trajectory starting from a point x:

$$\bar{f}(x) = \lim_{T \to \infty} \frac{1}{T} \int_0^T f(x_t)\, dt \quad (29)$$

The American mathematician G. D. Birkhoff (1884-1944) proved (this proposition is spoken of as the Birkhoff **ergodic theorem**) that mean value (29) exists everywhere except possibly for those x which form a set of measure zero or, as we say, **for almost all** values of x (**almost everywhere**). Earlier J. Neumann proved a weaker result (the **mean ergodic theorem**): the expression $\frac{1}{T}\int_0^T f(x_t)dt$ converges to $\bar{f}(x)$ in the mean square (i.e. in the sense of $L_2(M)$) for $T \to \infty$.

The function $\bar{f}(x)$ retains a constant value along every trajectory (why?) and is thus *invariant with respect to the shift along the trajectories*. Furthermore, it can be easily proved that

$$\int_{(M)} \bar{f}(x)\, d\mu = \int_{(M)} f(x)\, d\mu \quad (30)$$

Indeed, breaking (M) into small parts (ΔM_1), (ΔM_2), ..., (ΔM_q) and taking advantage of the invariance of the measure we obtain

$$\int_{(M)} f(x_t)\, d\mu \approx \sum_{k=1}^{q} f(x_{kt})\, \mu(\Delta M_k) = \sum_{k=1}^{q} f(x_{kt})\, \mu(\Delta M_k)_t \approx \int_{(M)} f(x)\, d\mu$$

(analyse these computations!) whence, after the passage to the limit, we derive the relation

$$\int\limits_{(M)} f(x_t)\,d\mu = \int\limits_{(M)} f(x)\,d\mu \tag{31}$$

Now we can write

$$\int\limits_{(M)} \left[\frac{1}{T}\int\limits_0^T f(x_t)\,dt\right] d\mu = \frac{1}{T}\int\limits_0^T dt \int\limits_{(M)} f(x_t)\,d\mu = \int\limits_{(M)} f(x)\,d\mu$$

which leads to (30) as $T \to \infty$.

Let us consider an important special case when the meaning of the function $\overline{f}(x)$ is particularly clear. Let (A) be a point set in (M) and the function $f_{(A)}(x)$ be defined as

$$f_{(A)}(x) = \begin{cases} 1 \text{ for } x \in (A) \\ 0 \text{ for } x \overline{\in} (A) \end{cases}$$

$f_{(A)}(x)$ is called the **characteristic function** of the set (A). We see that, by virtue of (29), the quantity $\overline{f}_{(A)}(x)$ is nothing but the **probability** of finding in (A) the moving point of the trajectory issued from the point x because it is equal to the average time spent by the moving point in the set (A).

A dynamic system is called **ergodic** if its every invariant function is equal to a constant (or, more precisely, is almost everywhere equal to a constant). For such a system we have $\overline{f}(x) \equiv \overline{f} = \text{const}$, and therefore it follows from (30) that

$$\overline{f} = \frac{1}{\mu(M)} \int\limits_{(M)} f(x)\,d\mu$$

that is $\overline{f}$ is equal to the mean value of the function $f(x)$ taken with respect to the invariant measure μ. Formula (29) thus takes the form

$$\lim_{T \to \infty} \frac{1}{T}\int\limits_0^T f(x_t)\,dt = \overline{f} = \frac{1}{\mu(M)} \int\limits_{(M)} f(x)\,d\mu$$

In other words, *for almost all x the time average of any function is equal to its space average.* This assertion plays an important role in physics (particularly, in statistical physics) because it provides some means for computing the density of probability (on this notion see e.g. [107], Sec. XVIII.8) for a system to be in a certain state by means of space averaging.

Ergodic systems possess an important feature known as **mixing property** which can readily be visualized. To this end, let us

interpret a given flow as a flow of a liquid in the phase space. Suppose that a "dye experiment" is carried out: a portion of the liquid occupying a volume (V) at moment t is dyed. We can pose the question as to in what way the dyed portion of the liquid is distributed in the space as $t \longrightarrow \infty$. Mathematically, this means that we are interested in the behaviour of the quantity $\mu((V)_t \cap (A))$ for $t \longrightarrow \infty$ where (A) is an arbitrary set and the symbol $\cap$ designates the operation of taking the common part (intersection) of sets. Indeed, the intersection $(V)_t \cap (A)$ is nothing but the portion of the dyed liquid found at time moment t in the mentally isolated domain (A). We have

$$\frac{1}{T}\int_0^T f_{(A)}(x_t)\,dt \xrightarrow[T\to\infty]{} \bar{f}_{(A)} = \frac{\mu(A)}{\mu(M)}$$

and, consequently, multiplying both sides by $f_{(V)}(x)$ and integrating with respect to x, we obtain

$$\frac{1}{T}\int_0^T \left[\int_{(M)} f_{(A)}(x_t)\, f_{(V)}(x)\,d\mu\right] dt \xrightarrow[T\to\infty]{} \frac{\mu(A)}{\mu(M)}\mu(V) \tag{32}$$

By virtue of formula (31), the integral in the square brackets is equal to

$$\int_{(M)} f_{(A)}(x)\, f_{(V)}(x_{-t})\,d\mu \tag{33}$$

But $f_{(V)}(x_{-t})$ is the characteristic function of the set $(V)_t$, and therefore integral (33) is equal to $\mu((V)_t \cap (A))$ (examine these considerations!). Substituting (33) into (32) we finally obtain

$$\frac{1}{T}\int_0^T \frac{\mu((V)_t \cap (A))}{\mu(A)}\,dt \xrightarrow[T\to\infty]{} \frac{\mu(V)}{\mu(M)}$$

Thus, the average percentage of the dyed liquid in (A) for large T is approximately equal to its percentage in (M). Under some additional requirements the mixing property can be proved in the "strong sense", that is as a stronger assertion

$$\frac{\mu((V)_t \cap (A))}{\mu(A)} \xrightarrow[t\to\infty]{} \frac{\mu(V)}{\mu(M)}$$

Generally, the mixing property means that *every part of the phase space is transferred into its every part uniformly in the long run.*

Now it seems expedient to ask how a given dynamic system can be tested for ergodicity. Unfortunately, at present there are no tests of this kind applicable to general systems. That is why in many physical problems this important property is introduced as

a hypothesis whose validity is justified by some qualitative analogies without rigorous proof.

To elucidate the difference between ergodic and nonergodic systems let us consider the following simple example. Take the autonomous system

$$\dot{\varphi}_1 = \omega_1, \quad \dot{\varphi}_2 = \omega_2 \tag{34}$$

where ω_1 and ω_2 are some constants and the coordinates φ_1 and φ_2 are interpreted as "angles" in the sense that the addition of any integers to them does not change the position of the representing point in the configuration manifold on which the corresponding flow is considered. In other words, system (34) is considered on a torus (see Sec. 11) or is represented on the entire

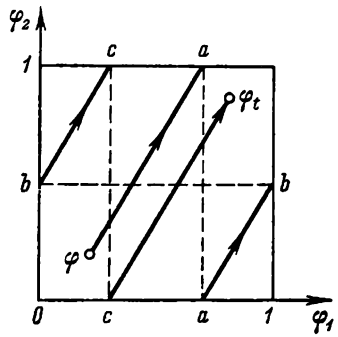


Fig. 161

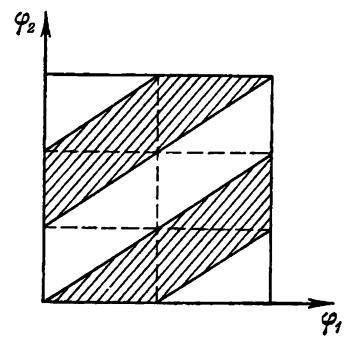


Fig. 162

$\varphi_1\varphi_2$-plane in which all the squares with unit sides and integral coordinates of the vertices are identified. The direct integration shows that the trajectories in the plane are straight lines (Fig. 161) and that the area is preserved under the shift along the trajectories. We thus deal with a dynamic system. But here there is a substantial difference between the cases when the "frequencies", ω_1 and ω_2, are commensurable $\left(\text{i.e. the quotient } \frac{\omega_1}{\omega_2} \text{ is rational}\right)$ and incommensurable $\left(\frac{\omega_1}{\omega_2} \text{ is irrational}\right)$. In the former case all the trajectories are closed, and it is easy to construct an invariant function not equal identically to a constant; for instance, if $\frac{\omega_1}{\omega_2} = \frac{3}{2}$ we can take as such a function the characteristic function of the region shaded in Fig. 162 (check it up!). Thus, this case is nonergodic.

In contrast to it, if the quotient $\frac{\omega_1}{\omega_2}$ is irrational all the trajectories are nonclosed. It can be shown that each trajectory, when

infinitely extended, is **everywhere dense** in the torus, that is passes arbitrarily close to its every point (although does not exhaust all the points of the torus; this resembles the properties of the set of all rational points of an interval which is everywhere dense in the interval but does not exhaust it). By the way, it follows that both α- and ω-limit sets of any trajectory coincide with the entire torus (more precisely, with its surface on which the flow is considered). We see that every piecewise continuous invariant function of the flow must be identically equal to a constant (why?). It turns out that this property remains valid when the condition of the piecewise continuity is dropped and consequently we deal with an ergodic system. (Let the reader verify that in this case the mixing property in the strong sense does not take place.)

It is easily proved that the set of the rational numbers on the straight line has measure zero, which means that irrational numbers are in a certain sense more typical than rational ones. Similarly, the ergodic case is more typical for this example than the nonergodic.

In conclusion we note that phase spaces dealt with in applications to conservative canonical systems (see Sec. VI.3.11) are usually noncompact but the whole construction described in this section can be used for the compact submanifold of the phase space corresponding to every given total-energy level.

Mathematical aspects of the *ergodic theory* can be found in [54] and [63].

§ 3. STABILITY OF SOLUTIONS

Stability in the sense of Lyapunov is a very important notion (e.g. see [107], Sec. XV.22) of theoretical and applied mathematics. Here we shall study this type of stability and some related questions. A more thorough treatment of the stability theory for solutions of differential equations can be found in [7], [19], [73], [78], and [92].

1. Introduction. Take a system of differential equations

$$\dot{y}_j = F_j(y_1, y_2, \ldots, y_n, t) \qquad (j = 1, 2, \ldots, n)$$

defined on an infinite time interval $t_0 \leqslant t < \infty$. As in § 2, we shall use the shortened form

$$\dot{\mathbf{y}} = \mathbf{F}(y, t) \tag{1}$$

Let $y = \tilde{y}(t)$ $(y = (y_1, \ldots, y_n))$ be a solution satisfying some initial conditions $y|_{t=t_0} = \tilde{y}_0$ which will be referred to as the **unperturbed solution** of (1). Every solution $y(t)$ of system (1) with any other initial data $t = t_0$, $y = y_0$ will be called a **perturbed solution.**

The unperturbed solution $\tilde{y}(t)$ is said to be **Lyapunov stable** if the perturbed solution corresponding to an arbitrary sufficiently small variation of the initial data is as close as desired to the unperturbed solution for all $t > t_0$. In other words, this means that

$$\max_{t_0 \leqslant t < \infty} |y(t) - \tilde{y}(t)| \to 0 \text{ for } |y_0 - \tilde{y}_0| \to 0 \qquad (2)$$

(Let the reader carefully examine this definition; it should be noted that it would be more precise to write "sup" instead of "max" since the maximum value may be achieved in the limit, at infinity.) If condition (2) is violated the unperturbed solution $\tilde{y}(t)$ is said to be **unstable** in the sense of Lyapunov.

There is a modification of the notion of stability introduced by Lyapunov: the unperturbed solution $\tilde{y}(t)$ is called **asymptotically stable** if it is stable (i.e. condition (2) is fulfilled) and, besides, for sufficiently small values of $|y_0 - \tilde{y}_0|$ the limiting relation $y(t) - \tilde{y}(t) \xrightarrow[t \to \infty]{} 0$ holds. There are a number of other types of stability which are of interest. For instance, consider the case when not only the initial data but the system of equations as well are perturbed so that the perturbed solution $y(t)$ satisfies the system of equations

$$\dot{y} = F(y, t) + \varphi(y, t)$$

Then the solution $\tilde{y}(t)$ is said to be **stable with respect to constant (constantly operating) perturbations** if

$$\max_{t_0 \leqslant t < \infty} |y(t) - \tilde{y}(t)| \to 0 \text{ for } |y_0 - \tilde{y}_0| \to 0 \text{ and } \max_{y, t} |\varphi(y, t)| \to 0$$

In real problems a perturbation of system (1) may be due to the application of some small external forces acting during the whole process of motion.

It can be proved that in some (most typical) cases stability with respect to perturbations is implied by asymptotic stability. For example, this is the case when the function F and the unperturbed solution are independent of time or are periodic in t with the same period.

A stability problem is usually reduced to investigating the stability of an equilibrium position by transferring the origin to the moving point, that is by means of the substitution $y = \tilde{y}(t) + x$. After this substitution has been made system (1) goes into

$$\dot{x} = \dot{y} - \dot{\tilde{y}}(t) = F(x + \tilde{y}(t), t) - \dot{\tilde{y}}(t)$$

or, in the shortened form,

$$\dot{\mathbf{x}} = \mathbf{f}(x, t) \tag{3}$$

and the unperturbed solution is transformed into the zero solution (a rest point at the origin). It should be remarked that even in the case of an autonomous original system (1) whose right-hand side is time-independent system (3) can nevertheless be nonautonomous since it involves $\tilde{y}(t)$. If the unperturbed solution is a rest point (that is $\tilde{y}(t) \equiv \text{const}$) of an autonomous system the transformed system (3) is also autonomous. It should also be taken into account that if the unperturbed solution $\tilde{y}(t)$ is periodic with period ω and system (1) is autonomous or periodic with period ω the resultant system (3) is also periodic with the same period.

In what follows we shall consider system (3) possessing a zero solution which will be taken as the unperturbed solution.

In some problems stability is interpreted not in the sense that all the components of the perturbed solution are small but as the smallness of some of them or of their certain combinations or, more generally, of some functionals dependent on the solution. For instance, we may be interested in whether the velocity of a system is small while the growth of the coordinates is inessential or in whether the integral of the squares of the deviations of the coordinates is small and the like. In such cases we speak about the **conditional stability** (relative to the chosen combinations or functionals, etc).

It is sometimes necessary to consider not all the possible perturbations of the initial data but only those subject to certain conditions; then we speak about the stability relative to the chosen class of perturbations.

2. First-Order Equations. The stability problem is readily solved, without resorting to the general theory, for differential equations of the first order. Let us begin with the equation

$$\dot{x} = f(x) \tag{4}$$

under the assumption that it possesses a solution identically equal to zero, that is $f(0) = 0$. The stability of this solution depends on the sign of $f(x)$ for the values of x close to zero.

Suppose that when x increases and passes through zero the function $f(x)$ changes its sign from $+$ to $-$. The isoclines of equation (4) being the straight lines $x = \text{const}$, the family of integral curves is located as shown in Fig. 163. The solution $x \equiv 0$ clearly is asymptotically stable in this case. The rate at which $x(t)$ tends to zero for $t \to \infty$ is easily determined since equation (4) is readily integrated; this rate depends on the rate at which $f(x)$ tends to zero as $x \to 0$. (Let the reader show that in the basic case when $f(x) \sim kx$, $k < 0$, for small x, the solution $x(t)$ tends

to zero at the rate of an exponential function while in the case $|f(x)| \sim |kx|^m$ $(m > l)$ only at the rate of a power function.)

For any other combinations of signs of $f(x)$ for small x the solution $x \equiv 0$ is unstable (why?).

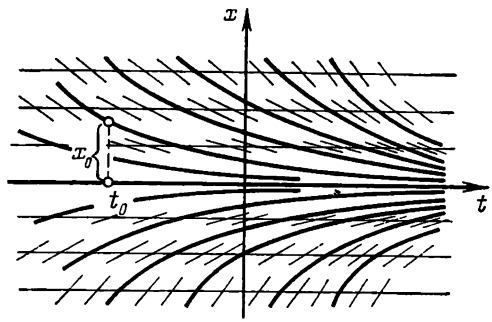


Fig. 163

Let us apply the above result to the equation

$$\dot{x} = f(x, \lambda)$$

involving the parameter λ where, for the sake of simplicity, the function f is assumed to be continuous. Suppose that the line (L) in the λx-plane determined by the equation $f = 0$ and the signs of f are as in Fig. 164. Then for every fixed value of λ the direction in which x varies is as shown by the arrows in the figure. This construction enables us to determine the parts of (L) to which the stable equilibrium positions correspond; these are shown in heavy lines in Fig. 164. We see that for $\lambda < \lambda_1$ and for $\lambda > \lambda_2$ there is only one stable stationary state, for $\lambda_1 < \lambda < \lambda_2$ there are two stable equilibrium states and one unstable state and the values $\lambda = \lambda_1$ and $\lambda = \lambda_2$ specify points of bifurcation. The existence of the interval $\lambda_1 < \lambda < \lambda_2$ on which x is a three-valued function of λ provides for an interesting phenomenon of jump-like changes of the solution when the parameter λ varies (Fig. 164). We see that if λ varies slowly (this condition of slowness makes it possible to consider the process as

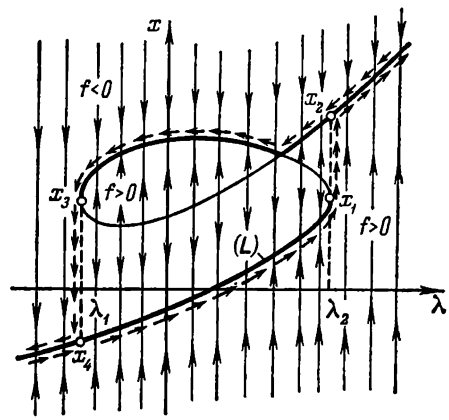


Fig. 164

being steady-state, stationary; such processes are said to be **quasi-static**) and increases to reach the values λ_2 the abscissa x of the equilibrium position continuously increases to reach x_1 after which it makes a jump, takes on the value x_2 and then proceeds to increase as λ grows. If λ varies in the opposite direction the abscissa x changes continuously until it attains the value x_3 after which it jumps to the value x_4. Thus, when the parameter λ varies in different directions it is necessary to use different branches of the line (L). This is known as **hysteresis phenomenon,** the region in Fig. 164 whose boundary passes through the points with ordinates x_1, x_2, x_3 and x_4 being called a **hysteresis loop (cycle).**

Now let us consider the equation

$$\dot{x} = f(x, t)$$

where $f(0, t) \equiv 0$, which means that the zero function is a solution of the equation. As before, assume that the sign of $f(x, t)$ is opposite to that of x for small $|x|$. Then the direction field is approximately the same as in Fig. 163, i.e. the zero solution is stable in this case as well. But if the slope of the field becomes too small as $t \to \infty$ the perturbed solutions do not necessarily tend to zero, that is there may be no asymptotic stability. Here is a typical example. Take the equation

$$\dot{x} = \varphi(t)\, x \qquad (\varphi(t) < 0)$$

Integrating we obtain

$$x = C \exp \int_{t_0}^{t} \varphi(t)\, dt \qquad (C = x(t_0))$$

Hence, if $\int_{t_0}^{\infty} \varphi(t)\, dt = -\infty$ the solution $x \equiv 0$ is asymptotically stable, and simply stable but not asymptotically if otherwise (examine this situation!).

3. Lyapunov's Function. Now let us proceed to study the many-dimensional autonomous system

$$\dot{\mathbf{x}} = \mathbf{f}(x) \tag{5}$$

where $\mathbf{f}(0) = 0$. The function $\mathbf{x} \equiv \mathbf{0}$ is thus a solution of (5), and we take it as the unperturbed solution. The *method of Lyapunov's function* widely used in the theory of stability and related sciences has an apparent mechanical analogue which we shall demonstrate by the example of a linear oscillator with one degree of freedom. The equation describing the oscillations is written in the form (e.g. see [107], (XV.115))

$$\ddot{x} + 2h\dot{x} + \omega_0^2 x = 0$$

Passing to the phase plane we reduce the equation to the system

$$\dot{x}_1 = x_2,\ \dot{x}_2 = -\omega_0^2 x_1 - 2hx_2 \quad (x_1 = x,\ x_2 = \dot{x}) \tag{6}$$

As can be easily verified, the total energy of the oscillator is equal, to within a proportionality factor, to the expression

$$E(x_1,\ x_2) = \omega_0^2 x_1^2 + x_2^2 \tag{7}$$

(e.g. see [107], formula (XV.72)) and therefore has a minimum value at the rest point $x_1 = x_2 = 0$; the lines $E = \text{const}$ in the phase plane are depicted in Fig. 165. If $h > 0$, that is if there is a positive friction, the total energy decreases in the process of oscillations; this property can aslo be proved formally:

$$\dot{E} = \omega_0^2 2x_1\dot{x}_1 + 2x_2\dot{x}_2 = 2\omega_0^2 x_1 x_2 + 2x_2(-\omega_0^2 x_1 - 2hx_2) = -4hx_2^2 < 0 \tag{8}$$

We see that the closed lines of constant total energy in the phase plane are intersected by the trajectories from their outer side to the interior whence it follows that the equilibrium state $x_1 = x_2 = 0$ is asymptotically stable. If $h = 0$, which indicates the absence of the friction, the total energy of motion is constant and the motion in the phase plane is in the ellipses shown in Fig. 165, the equilibrium state thus being nonasymptotically stable. Finally, if $h < 0$ (this is the case of negative friction which is realizable for systems having an external source of energy) we have $\dot{E} > 0$, that is the ellipses are intersected by the trajectories from the interior to the exterior and the rest point $x_1 = x_2 = 0$ is unstable.

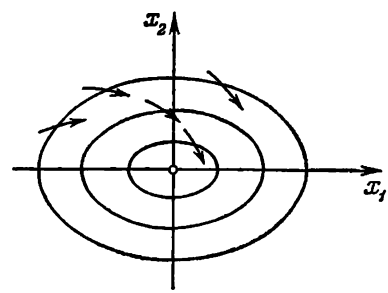


Fig. 165

In this argument the fact that E is proportional to the total energy is not very important; here the essential feature is that the function $E(x_1,\ x_2)$ behaves in a certain manner in the neighbourhood of the equilibrium point (namely, has a minimum), and the stability (or instability) is specified by the character of the variation of this function along the trajectories (that is depends on whether E decreases or increases). This construction can be generalized to obtain, for the general system (5).

Lyapunov's Theorem on Stability. *Let in a neighbourhood of the point* $x = 0$ *there exist a continuous function* $V(x)$ (referred to as the **Lyapunov function**) *satisfying the conditions that* $V(0) = 0$ *and* $V(x) > 0$ *for* $x \neq 0$ (by analogy with quadratic forms, such a

function is said to be **positive definite**) *and let the inequality*

$$\dot{V} \equiv \operatorname{grad} V(x) \cdot \mathbf{f}(x) \leqslant 0 \tag{9}$$

be fulfilled. Then the unperturbed solution $x \equiv 0$ *is stable.*

To prove this apparent proposition, suppose that we are given a small number $\varepsilon > 0$. Then the function $V(x)$ has a positive minimum value $m_\varepsilon > 0$ on the sphere $(S)_\varepsilon$ with centre at $x = 0$ and radius ε since it is continuous and positive on that sphere. Now let us choose a sphere $(S)_\delta$ of sufficiently small radius $\delta > 0$ such that $V(x) < m_\varepsilon$ inside $(S)_\delta$. Then, if $|\mathbf{x}_0| < \delta$, i.e. if the initial point of the trajectory lies in the interior of $(S)_\delta$, condition (9) implies that the whole trajectory always remains inside $(S)_\varepsilon$ for $t > t_0$ (why?), which means that the solution $x \equiv 0$ is stable.

Of course, in this theorem we can as well use a function $V(x)$ subject to the condition that $V(0) = 0$, $V(x) > 0$ for $x \neq 0$ and $\dot{V}(x) \geqslant 0$ (this remark also refers to the theorems below).

The above argument can be modified (but here we do not go into particulars) to prove that *if condition* (9) *is replaced by the stronger requirement that the function* $\dot{V}(x)$ *should be negative definite the solution* $x \equiv 0$ *is asymptotically stable.* Condition (9) does not guarantee asymptotic stability because it does not exclude the case when $\dot{V} \equiv 0$ along some trajectories, that is $V = \text{const}$, which means that such trajectories do not necessarily enter the point $x = 0$ for $t \to \infty$. But *if*, in addition to (9), we require that *the manifold on which* $\dot{V} = 0$ *should not entirely contain, for* $x \neq 0$, *any arc of a trajectory* the above phenomenon cannot take place and *the solution* $x \equiv 0$ *is asymptotically stable.* In particular, this useful comment applies to the example of the linear oscillator (formulas (6)-(8)) in the case $h > 0$: the derivative $\dot{E}$ is not negative definite in this example but it vanishes only on the straight line $x_2 = 0$ which does not contain any entire arc of a trajectory (why?).

If Lyapunov's function $V(x)$ satisfying the stability conditions is known in a wider region than a small neighbourhood of the point $x = 0$ it can be used for determining a part of the domain of attraction of this stable solution. For instance, suppose that the function $V(x)$ is defined throughout the x-space and that the manifold on which $\dot{V} = 0$ does not contain, for x where $\mathbf{f}(x) \neq 0$, any arc of a trajectory. Then, if $V(x)$ has a minimum at the point $x = 0$ the corresponding "pit" (see Sec. VI.3.13) is entirely contained in the domain where the solution $x \equiv 0$ attracts the trajectories although does not necessarily exhaust this domain. But if, in addition, it is known that $x = 0$ is a single rest point of system (5) and $V(\infty) = \infty$, all the trajectories tend to the rest point $x = 0$ for

$t \to \infty$. In this case we say that the solution $x \equiv 0$ is **asymptotically stable in the large** (possesses **global asymptotic stability**).

For the solution $x \equiv 0$ to be unstable it is sufficient that there exist a function $V(x)$ in a neighbourhood of the point $x = 0$ which assumes positive values at points lying arbitrarily close to $x = 0$ and whose time-derivative along the trajectories $\dot{V}(x)$ is positive in all regions where $V(x) > 0$. Indeed, for definiteness, let the function V be positive in the region shaded in Fig. 166 where (S) is a circle of a small fixed radius. Then for any initial point x_0 belonging to this region and lying arbitrarily close to the origin the values of $V(x)$ increase along the trajectory issued from that point, and therefore the moving point of such a trajectory cannot pass arbitrarily close to the boundary of the shaded domain (on which $V = 0$). Hence, $\min \dot{V} > 0$ along the whole trajectory, and consequently the moving point must reach (S) in a finite time, which indicates the instability (let the reader carefully analyse this argument). This theorem generalizes similar results obtained by Lyapunov; it was established in 1934 by N. G. Chetaev (1902-1959), a Soviet scientist.

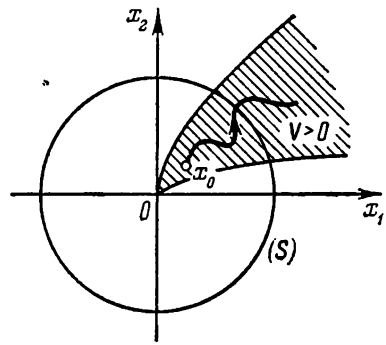


Fig. 166

The theorem also remains true if we assume, as above, that $\dot{V}$ may vanish at some points belonging to the region where $V > 0$ but the manifold of all such points does not contain, for $x \neq 0$, any entire arc of a trajectory. This useful remark and many other generalizations and converses of Lyapunov's theorems are due to N. N. Krasovsky (born in 1924), a Soviet mathematician and mechanician (see [73]).

The theorems given here can be generalized to nonautonomous systems (3) and also to Lyapunov's functions of the form $V(x, t)$ if some corrections are inserted in their formulations and proofs. For instance, in Lyapunov's basic stability theorem we must speak of a positive definite function $V(x, t)$ in the sense that $V(x, t) \geqslant V_1(x)$ where the function $V_1(x)$ is positive definite in the ordinary sense (for example, the function $|\mathbf{x}|(1+t^2)^{-1}$ is positive but not positive definite). The derivative $\dot{V}$ is now computed by the formula $\dot{V}(x, t) = V'_t + \operatorname{grad} V(x, t) \cdot \mathbf{f}(x, t)$. The formulation of the theorem on asymptotic stability must involve the additional condition

$$\max_t V(x, t) \xrightarrow[x \to 0]{} 0$$

(which, for instance, excludes such functions as $t|x|$). Finally, in the theorem on instability it is necessary to require additionally that in every domain where V is greater than a positive constant the function $\dot{V}$ should also exceed a positive constant.

When the functions involved depend on t it may turn out that the conditions of Lyapunov's theorems are only fulfilled for $t \geqslant t_1$ where $t_1 > t_0$. In this connection it should be taken into account that such properties as stability, instability and asymptotic stability (which are local with respect to x) are independent of the time moment at which the initial conditions are set because for small $|\mathbf{x}(t_0)|$ the quantity $|\mathbf{x}(t_1)|$ is also small and vice versa. But the properties that characterize the system in the large, such as the global asymptotic stability, may depend on the initial value of time t (let the reader analyse these facts!).

The method of Lyapunov's functions is usually referred to as **Lyapunov's second method** in the theory of stability (**Lyapunov's first method** is understood as one based on constructing solutions in a direct manner in the form of a sum of a series and the like). Generally, the methods of studying the properties of solutions without resorting to their exact analytic expressions or numerical computation belong to the so-called **qualitative theory of differential equations.**

Some examples demonstrating applications of Lyapunov's theorems will be discussed later.

In recent years much attention has been paid to the investigation of equations possessing the property of global (total) stability in the sense that all their solutions are stable. Below is a typical result of this kind. Let the right-hand side of system (3) satisfy the inequality

$$(\mathbf{f}(x, t) - \mathbf{f}(y, t)) \cdot (\mathbf{x} - \mathbf{y}) < 0 \quad (x \neq y)$$

for all the values of the arguments. Then any two different solutions $\mathbf{x}(t)$ and $\mathbf{y}(t)$ of the system fulfil the relation

$$\frac{d}{dt}(|\mathbf{x}-\mathbf{y}|^2) = \frac{d}{dt}[(\mathbf{x}-\mathbf{y})\cdot(\mathbf{x}-\mathbf{y})] = 2(\mathbf{f}(x, t)-\mathbf{f}(y, t))\cdot(\mathbf{x}-\mathbf{y}) < 0$$

which means that these solutions tend to each other as t increases. On these questions see [68].

4. Stability in Linear Approximation. When testing the zero solution of equation (3) for stability we only deal with small values of $\mathbf{x}$ and therefore it appears natural to expand the right-hand side in powers of x_j, which results in

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \boldsymbol{\varphi}(x, t) \left(\mathbf{A}(t) = \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\right)_{\mathbf{x}=0}\right) \tag{10}$$

where $\boldsymbol{\varphi}$ is the combination of all the terms of order higher than

the first relative to x. Now, discarding the higher-order terms, we arrive at the linearized system

$$\dot{\mathbf{x}} = \mathbf{A}(t)\,\mathbf{x} \tag{11}$$

referred to as the **linear approximation** for (3) ((10)) or the **system of equations of first approximation.**

For the sake of simplicity, we shall suppose that $\varphi(x, t) = o(|\mathbf{x}|)$ in equation (10), that is φ has the order of smallness higher than the first relative to $|\mathbf{x}|$ uniformly with respect to t. (This condition excludes such terms as tx_j^2 whose coefficients in higher powers of x_j increase indefinitely as $t \longrightarrow 0$.) If the zero solution of system (10) is stable for all the additional terms φ of that type the solution $\mathbf{x} \equiv \mathbf{0}$ of system (3) is said to be **stable in linear (first) approximation** (the definition of instability in linear approximation is formulated in like manner).

System (11) being linear, it follows from formula (1.10) that *for the zero solution of system* (11) *to be stable it is necessary and sufficient that its all solutions be bounded for* $t \longrightarrow \infty$; *for asymptotic stability it is necessary and sufficient that all the solutions tend to zero as* $t \longrightarrow \infty$ (prove it!). Note that for linear systems, in contrast to general (nonlinear) ones, the stability of one of the solutions implies the stability of all the other solutions, and therefore, in connection with (11), we can simply speak of a stable system of linear equations; the same refers to asymptotic stability and to instability.

Let us consider a simpler case when $\mathbf{A}$ is a constant matrix (independent of t). According to Sec. 1.1, *for system* (11) *with constant matrix* $\mathbf{A}$ *to be asymptotically stable it is necessary and sufficient that all the roots of the characteristic equation*

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0 \tag{12}$$

have negative real parts. (The rate at which the general solution tends to zero for $t \longrightarrow \infty$, the so-called **degree of stability**, is determined by the greatest of these real parts.)

It turns out that in this case *the solution* $\mathbf{x} \equiv 0$ *of system* (10) *is asymptotically stable* as well.

Here we do not present the general proof and limit ourselves to the case when all the roots of equation (12) are real and distinct. The substitution $\mathbf{x} = \mathbf{H}\mathbf{y}$ reduces the system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ to $\mathbf{H}\dot{\mathbf{y}} = \mathbf{A}\mathbf{H}\mathbf{y}$, that is to $\dot{\mathbf{y}} = \mathbf{H}^{-1}\mathbf{A}\mathbf{H}\mathbf{y}$, and therefore, under the hypothesis assumed, it is always possible to take a nondegenerate matrix $\mathbf{H}$ such that the coefficient matrix of the transformed system takes the diagonal form diag $(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Suppose that this transformation has been already performed and that the system of form (11)

we deal with has a diagonal coefficient matrix. Then we can put

$$V(x)=\sum_j|\lambda_j|x_j^2=-(\mathbf{Ax},\ \mathbf{x})$$

(remember that $\lambda_j<0$, $j=1, 2, \ldots, n$). This function is obviously positive definite and its total time-derivative along the integral curves of system (10) is expressed (check it up!) as

$$\dot{V}=-2(\mathbf{A}\dot{\mathbf{x}},\ \mathbf{x})=-2(\mathbf{A}^2\mathbf{x},\ \mathbf{x})+o(|\mathbf{x}|^2)=-2\sum_j\lambda_j^2x_j^2+o(|\mathbf{x}|^2)$$

Thus, the conditions of Lyapunov's theorem hold, which is what we set out to prove.

In the case of imaginary roots of the characteristic equation it is often expedient to apply linear transformations leading to complex unknown functions. For instance, the system of equations

$$\dot{x}_1=\mu x_1+\nu x_2,\ \dot{x}_2+-\nu x_1+\mu x_2,\ \dot{x}_3=\lambda x_3+\varphi(x_1) \qquad (13)$$

is reduced, by means of the substitution

$$z_1=x_1+ix_2,\ z_2=x_1-ix_2,\ z_3=x_3$$

or, equivalently, by the change

$$x_1=\frac{z_1+z_2}{2},\quad x_2=\frac{z_1-z_2}{2i},\quad x_3=z_3 \qquad (14)$$

to the system

$$\dot{z}_1=(\mu-i\nu)z_1,\ \dot{z}_2=(\mu+i\nu)z_2,\ \dot{z}_3=\lambda z_3+\varphi\left(\frac{z_1+z_2}{2}\right) \qquad (15)$$

(check up the calculations!) where the functions $z_j(t)$ are no longer real and can assume complex values. There are two approaches to a system of type (15). In the first approach we regard z_j as arbitrary complex variables, and then the Lyapunov function $V(z)$ must be defined for complex values of the arguments and take on real values for all such z. For instance, it can be constructed as the Hermitian quadratic form

$$V=\mu(z_1z_1^*+z_2z_2^*)+\lambda z_3z_3^*$$

and then

$$\dot{V}=2\mu^2(z_1z_1^*+z_2z_2^*)+2\lambda^2z_3z_3^*+\lambda(z_3^*\varphi+z_3\varphi^*)$$

It is clear that after the stability or the asymptotic stability of the zero solution of system (15) has been established the result obtained applies to system (13) as well. But in the case of instability of system (15) some additional considerations are involved because there may be no real solutions of the original system (13) corresponding to complex solutions of (15) going away from the point $z=0$. In the second approach we only consider the values

of z to which there correspond real values of x. Formulas (14) then indicate that z_1 and z_2 must be complex conjugate and z_3 must be real. In particular, it is necessary to resort to this approach if the function φ is defined only for real values of its arguments. In the second approach the Lyapunov function must be real for all these values of z but may assume imaginary values for the other z's; we can also regard it as being defined solely for the above-indicated values of z, and then the notions of stability or instability for the original and transformed systems coincide. It should also be taken into account that if to the given initial values of z there correspond real values of x then the whole solution $z(t)$ of system (15) possesses this property (why?).

The above construction of the Lyapunov function is impractical since the reduction of a matrix to the diagonal form is connected with the solution of equation (12). Therefore it is sometimes advisable to use the following technique. Let $\mathbf{C}$ be an arbitrary symmetric matrix. Then it is possible to choose a symmetric matrix $\mathbf{U}$ such that

$$\frac{d}{dt}(\mathbf{Ux}, \mathbf{x}) \equiv (\mathbf{Cx}, \mathbf{x}) \tag{16}$$

(the derivative is taken along the trajectories of the system $\dot{\mathbf{x}} = \mathbf{Ax}$ which is supposed to be asymptotically stable), the construction of $\mathbf{U}$ reducing to the solution of a system of linear algebraic equations. Indeed, equality (16) can be rewritten in the form

$$2(\mathbf{UAx}, \mathbf{x}) \equiv (\mathbf{Cx}, \mathbf{x}) \tag{17}$$

Passing in (17) to scalar equations and equating the coefficients in the same combinations $x_j x_k$ we obtain a system of $N = \frac{n(n+1)}{2}$ equations of the first degree with N unknowns (which are the distinct elements of the matrix $\mathbf{U}$; why is their number equal to $N = \frac{n(n+1)}{2}$?). This system is solvable if and only if the corresponding homogeneous system possesses solely the trivial solution. But this condition is in fact fulfilled because if for $\mathbf{C} = \mathbf{0}$ there existed a solution $\mathbf{U} \neq \mathbf{0}$, equation (16) would imply that $(\mathbf{Ux}, \mathbf{x}) = \text{const}$ for any solution $\mathbf{x}(t)$, which contradicts the asymptotic stability (why?).

Thus, we have proved that the matrix $\mathbf{U}$ can in fact be constructed (incidentally, we have proved its uniqueness!). The technique we are going to present consists in the following. Take a positive definite symmetric matrix as the matrix $\mathbf{C}$ entering into the above argument. Then the asymptotic stability of the system $\dot{\mathbf{x}} = \mathbf{Ax}$ implies that the quadratic form $(\mathbf{Ux}, \mathbf{x})$ is negative definite. Indeed,

by virtue of the theorem on instability given in Sec. 3, the function $V(x)=(\mathbf{Ux}, \mathbf{x})$ cannot assume positive values. This means that then the relation $V(x)=0$ is impossible for $x\neq 0$ since $\dot{V}(x)>0$. Hence, the function $V(x)$ can be taken as the Lyapunov function.

By the way, we have $(\mathbf{UAx}, \mathbf{x})=(\mathbf{A^*Ux}, \mathbf{x})$ (why?), and consequently equality (17) can be rewritten in the form $((\mathbf{Ua}+\mathbf{A^*U})\mathbf{x}, \mathbf{x})=(\mathbf{Cx}, \mathbf{x})$. But since the matrix $\mathbf{UA}+\mathbf{A^*U}$ is symmetric, this relation is equivalent to the equality

$$\mathbf{UA}+\mathbf{A^*U}=\mathbf{C} \tag{18}$$

We have thus proved that matrix equation (18) possesses a uniquely determined symmetric solution $\mathbf{U}$ for any matrix $\mathbf{A}$ with negative real parts of its eigenvalues and any symmetric matrix $\mathbf{C}$. In particular if the matrix $\mathbf{C}$ is positive definite (that is the quadratic form $(\mathbf{Cx}, \mathbf{x})$ is positive definite or, which is the same, the eigenvalues of $\mathbf{C}$ are positive) the matrix $\mathbf{U}$ is negative definite and vice versa.

The possibility of constructing the above-described quadratic form $(\mathbf{Ux}, \mathbf{x})$ proved here is a particular case of Lyapunov's general theorem asserting that the equation $\frac{d}{dt}U_m(x)=C_m(x)$ where $C_m(x)$ is a given multilinear form of degree m and $U_m(x)$ the sought-for form of the same degree is solvable if and only if none of the sums of m eigenvalues (some of them can be involved repeatedly) of the matrix $\mathbf{A}$ is equal to zero $\left($the symbol $\frac{d}{dt}$ designates the differentiation along the trajectories of the system $\dot{\mathbf{x}}=\mathbf{Ax}\right)$.

The above construction of the matrix $\mathbf{U}$ is, in particular, applied to testing linear systems, perturbed by small nonlinear terms, for total asymptotic stability (asymtotic stability in the large). Let us consider the system

$$\dot{\mathbf{x}}=\mathbf{Ax}+\boldsymbol{\psi}(x)\mathbf{x} \tag{19}$$

under the assumption that the system $\dot{\mathbf{x}}=\mathbf{Ax}$ is asymptotically stable. Taking a negative definite quadratic form $(\mathbf{Cx}, \mathbf{x})$ and constructing the positive definite form $(\mathbf{Ux}, \mathbf{x})$ whose derivative along the trajectories of the system $\dot{\mathbf{x}}=\mathbf{Ax}$ is equal to $(\mathbf{Cx}, \mathbf{x})$ we see that its derivative along the trajectories of system (19) is expressed by the formula

$$\frac{d}{dt}(\mathbf{Ux}, \mathbf{x})=(\mathbf{Cx}, \mathbf{x})+2(\mathbf{Ux}, \boldsymbol{\psi}(x)\mathbf{x})$$

If for any fixed value of $\boldsymbol{\psi}$ the right-hand side of the latter relation (in which x is thus regarded as a variable quantity in all the

terms except the expression $\psi(x)$ which is replaced by a constant) is a positive definite quadratic form (this can be tested by the Sylvester's theorem), the zero solution of system (19) is totally (globally) asymptotically stable.

It is similarly proved that *if at least one root of equation* (12) *has a positive real part the solution* $\mathbf{x} \equiv \mathbf{0}$ *of system* (10) *is unstable.*

The cases when equation (12) possesses no roots with positive real parts but has at least one root with zero real part (i.e. a pure imaginary root) are termed **critical**; they involve a more sophisticated investigation. In critical cases the necessary and sufficient condition for stability of a linear system is expressed in terms of the properties of the Jordan submatrices (see Sec. 1.1) of its coefficient matrix while the stability of nonlinear system (10) may essentially depend on the structure of the function φ (Sec. 5).

The signs of the real parts of the roots of equation (12) can be determined with the aid of the methods given in Sec. II.3.10.

The case of a periodic matrix $\mathbf{A}(t)$ admits of a similar treatment. Namely, it turns out that *if the moduli of all the multipliers of periodic system* (11) *are less than unity the solution* $\mathbf{x} = \mathbf{0}$ *of periodic system* (10) *is asymptotically stable in linear approximation; if at least one of the moduli exceeds unity the solution* $\mathbf{x} = \mathbf{0}$ *is unstable.* This proposition immediately follows from the possibility of reducing a linear periodic system to a system with constant coefficients (see Sec. 1.2). To test a periodic system for asymptotic stability it is sometimes expedient to apply the methods of Sec. II.3.10 to the logarithm of the monodromy matrix (Sec. 1.2). But, on the other hand, since the monodromy matrix is usually found by means of numerical integration, in concrete examples the stability or the instability of periodic system (11) can be established in a sufficiently effective way without resorting to the theory, by carrying out numerical integration of the system over several periods for randomly chosen initial data. The increase or decrease of the absolute value of the solution is determined by the greatest in its modulus multiplier; to be sure of the validity of the conclusion on the stability drawn from the results of the numerical integration it is advisable to test a number of variants of initial data.

Unfortunately, for coefficient matrices $\mathbf{A}(t)$ of a more complicated nature (compared with those considered here and in Secs. 1.8 and 9) there are at present no effective tests for stability. Therefore, in concrete problems we sometimes perform a computational experiment with the aid of numerical integration of the given system for randomly chosen initial data on an interval whose length is sufficient for judging upon stability or instability. (The absolute value of every nonzero solution of the scalar equ-

ation $\dot{x}=ax$ increases or decreases 10 times on the interval $\Delta t = \frac{\ln 10}{|a|}=\frac{2.3}{|a|}$, which indicates that in the general case the length of the interval of integration should be at least several times the reciprocal of the greatest in its modulus coefficient of the system.) To watch the behaviour of the solution it is convenient to compute the value of $\frac{1}{t}\ln|\mathbf{x}(t)|$ in the process of the numerical integration. If the solution of system (11) reveals exponential damping or increases exponentially the solution $\mathbf{x}\equiv\mathbf{0}$ of system (10) is, respectively, asymptotically stable or unstable.

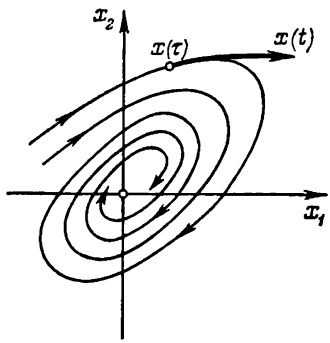


Fig. 167

When investigating a system of form (11) it is sometimes allowable to replace it by the system with constant coefficients $\dot{\mathbf{x}}=\mathbf{A}(\tau)\mathbf{x}$ considered for arbitrary fixed values of the parameter, but the reader should be aware that in the general case the asymptotic stability of the latter system is neither necessary nor sufficient for the asymtotic stability of system (11). This means that such a replacement requires some additional conditions (which we shall not discuss here). To elucidate what has been said, let us consider the case $n=2$. Suppose that for every fixed τ the system $\dot{\mathbf{x}}=\mathbf{A}(\tau)\mathbf{x}$ has a stable nodal point at the origin (see Fig. 167) and that when τ is varied the family of the trajectories turns, as a whole, about the point x 0. Then, although every small arc of the trajectory $x(t)$ of system (11) is close (in the vicinity of the point $t=\tau$) to the corresponding small arc of the trajectory of the system $\dot{\mathbf{x}}=\mathbf{A}(\tau)\mathbf{x}$, to describe the whole trajectory of system (11) we must pass to the turned trajectories of the system $\dot{\mathbf{x}}=\mathbf{A}(\tau)\mathbf{x}$ for subsequent time moments, which may result, as is shown in Fig. 167, in the increase of the deviation of the trajectories of the systems from the origin, which leads to instability. Hence, as has been already mentioned, the possibility of the application of this technique involves some additional restrictions.

5. Critical Cases. In Sec. 4 we have mentioned critical cases in which the stability of the zero solution of system (10) cannot be established on the basis of the analysis of the system of the first approximation (11) with a constant or periodic matrix $\mathbf{A}(t)$. A. M. Lyapunov carried out a thorough investigation of the case of one zero or two pure imaginary complex conjugate roots of the

characteristic equation. Since then many other cases were studied but these questions involve sophisticated techniques which we shall not discuss here at length. As an instance, we shall consider the simplest case of one zero root for the case $n=2$.

Take the system of equations

$$\dot{x}=ax+by+\varphi(x, y), \quad \dot{y}=cx+dy+\psi(x, y) \tag{20}$$

where the expansions of the functions φ and ψ in powers of x_1 and y_1 contain only terms of order higher than the first. Suppose that the characteristic equation of the corresponding system of linear approximation has one zero root and one negative root. Then, as in Sec. 4, we can pass to the system

$$\dot{x}_1=\varphi_1(x_1, y_1), \quad \dot{y}_1=\lambda y_1+\psi_1(x_1, y_1) \ (\lambda<0) \tag{21}$$

To investigate system (21) it is necessary to make one more transformation. Since the equation

$$\lambda y_1+\psi_1(x_1, y_1)=0 \tag{22}$$

satisfies the conditions of the theorem on the existence of the implicit function (see equations (II.2.26) and the corresponding discussion), it is possible to determine from it the function $y_1=f(x_1)$ in the form of a sum of a series without linear terms (why?). Putting

$$\xi=x_1, \quad \eta=y_1-f(x_1) \tag{23}$$

we obtain from (21) the system

$$\dot{\xi}=\Phi(\xi, \eta), \quad \dot{\eta}=\lambda\eta+\Psi(\xi, \eta) \tag{24}$$

(check it up!) where

$$\Phi(\xi, \eta)=\varphi_1(\xi, \eta+f(\xi))$$
$$\Psi(\xi, \eta)=\psi_1(\xi, \eta+f(\xi))-\psi_1(\xi, f(\xi))-f'(\xi)\varphi_1(\xi, \eta+f(\xi)) \tag{25}$$

Expanding these functions as

$$\left.\begin{aligned}\Phi(\xi, \eta)&=\Phi_0(\xi)+\Phi_1(\xi)\eta+\Phi_2(\xi)\eta^2+\dots\\ \Psi(\xi, \eta)&=\Psi_0(\xi)+\Psi_1(\xi)\eta+\Psi_2(\xi)\eta^2+\dots\end{aligned}\right\} \tag{26}$$

we see from (25) that the expansion of $\Psi_0(\xi)=\Psi(\xi, 0)$ in powers of ξ begins with terms of higher order than the expansion of $\Phi_0(\xi)$ (the change of variables (23) has been introduced to achieve this). As is shown below, the character of stability of the zero solution of system (24) (and therefore of system (20) as well) is determined by the structure of the first term of the expansion of the function $\Phi_0(\xi)$.

Let us begin with the case

$$\Phi_0(\xi)=\alpha\xi^2+\dots \qquad (\alpha\neq 0)$$

where the dots designate the terms of higher order, and choose Lyapunov's function of the form

$$V(\xi, \eta) = \xi + k\xi\eta + l\eta^2 \tag{27}$$

with the undermined coefficients k and l which will be specified later. According to (26), we have

$$\Phi(\xi, \eta) = \alpha\xi^2 + \beta\xi\eta + \gamma\eta^2 + \dots$$

and, consequently, (24) and (27) imply that

$$\dot{V} = \alpha\xi^2 + \beta\xi\eta + \gamma\eta^2 + \lambda k\xi\eta + 2\lambda l\eta^2 + \dots$$

Hence, if we put $k = -\frac{\beta}{\lambda}$ and $l = \frac{\alpha - \gamma}{2\lambda}$, the time-derivative of V is expessed, to within the terms of higher order, as $\dot{V} = \alpha(\xi^2 + \eta^2)$. Now we can apply the theorem on instability given in Sec. 3 whose all conditions (possibly to within the sign of the function V) are fulfilled, and thus the zero solution of system (20) is unstable in the case under investigation.

The same result can also be proved (by means of a more complicated construction) in the general case when the expansion of the function $\Phi_0(\xi)$ begins with an arbitrary even power of ξ.

Suppose now that the expansion of the function $\Phi_0(\xi)$ begins with the third degree of ξ. Then we can write

$$\Phi(\xi, \eta) = (\alpha\xi^3 + \dots) + (\beta\xi + \gamma\xi^2 + \dots)\eta + \dots \qquad (\alpha \neq 0)$$
$$\Psi(\xi, \eta) = (\delta\xi^4 + \dots) + (\varepsilon\xi + \dots)\eta + \dots$$

Let us take Lyapunov's function of the form

$$V = \frac{1}{2}\xi^2 + k_1\xi^2\eta + k_2\xi^3\eta + l\eta^2 \tag{28}$$

Computing the derivative $\dot{V}$ and writing down only the terms of order not higher than the fourth relative to ξ and not higher than the second relative to η we obtain (check it up!)

$$\dot{V} = \alpha\xi^4 + \beta\xi^2\eta + \gamma\xi^3\eta + k_1\lambda\xi^2\eta + k_1\varepsilon\xi^3\eta + k_2\lambda\xi^3\eta + 2l\lambda\eta^2 + o(\xi^4) + o(\eta^2)$$

Hence, if we put $k_1 = -\frac{\beta}{\lambda}$, $k_2 = -\frac{\gamma + k_1\varepsilon}{\lambda}$ and $l = \frac{\alpha}{2\lambda}$ the derivative is written as

$$\dot{V} = \alpha(\xi^4 + \eta^2) + o(\xi^4) + o(\eta^2) \tag{29}$$

Consequently, if $\alpha < 0$, then, by virtue of Sec. 3, the zero solution of system (20) is asymptotically stable, and if $\alpha > 0$ it is unstable. The same result is obtained when the expansion of the function $\Phi_0(\xi)$ begins with any odd power of ξ.

Finally, let $\Phi_0(\xi) \equiv 0$. It then follows that $\Psi_0(\xi) \equiv 0$, and system (24) takes the form

$$\dot{\xi} = \Phi_1(\xi)\eta + \Phi_2(\xi)\eta^2 + \ldots, \quad \dot{\eta} = \lambda\eta + \Psi_1(\xi)\eta + \Psi_2(\xi)\eta^2 + \ldots$$
$$(\Phi_1(0) = \Psi_1(0) = 0) \qquad (30)$$

This system is satisfied if we put $\xi = \text{const}$, $\eta = 0$; hence we obtain the straight line $\eta = 0$ in the ξ η-plane whose every point is stationary. For $\eta \neq 0$ we derive from (30) the equation

$$\frac{d\xi}{d\eta} = \frac{\Phi_1(\xi) + \Phi_2(\xi)\eta + \ldots}{\lambda + \Psi_1(\xi) + \Psi_2(\xi)\eta + \ldots}$$
$$(\xi = \xi(\eta)) \qquad (31)$$

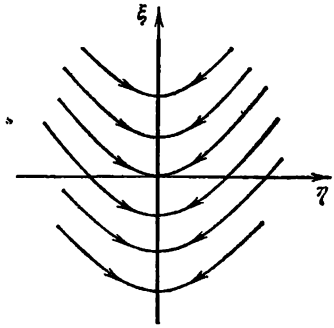


Fig. 168

for which the origin serves as a regular (not singular) point. Hence, the trajectories of system (30) go along the integral curves of equation (31). The second equation (30) indicates that the direction of motion along the trajectories is as shown in Fig. 168. Therefore, in this case the zero solution is stable but not asymptotically stable. The ξ-axis (whose equation is (22)) is an inertia zone (see Sec. 2.10) of the system in question.

The investigation of critical cases has an important application to testing the behaviour of solutions on the boundary of the stability region of a system involving parameters. For instance, let us consider a system of form (20) involving one parameter:

$$\left.\begin{aligned} \dot{x} &= a(s)x + b(s)y + \varphi(x, y, s) \\ \dot{y} &= c(s)x + d(s)y + \psi(x, y, s) \end{aligned}\right\}$$

(the case of an arbitrary number of parameters and unknown functions is treated similarly). Suppose that when the parameter s increases and passes through a point s_0 one of the two negative roots of the characteristic equation becomes positive; then the stability region (in the space of parameters, that is in this case on the s-axis) lies to the left of the point s_0: $s < s_0$. If the coefficients of the system continuously depend on the parameter the roots are also continuous in s, and thus, for $s = s_0$, we just have the case of one zero root studied above.

Suppose that for $s = s_0$ the zero solution is asymptotically stable. As was shown (in the example with $\Phi_0(\xi) \sim \alpha\xi^3$), in this case it is possible to construct a positive definite Lyapunov function (28) whose total time-derivative (29) is negative definite, which guarantees the asymptotic stability. Let us now investigate what

happens if we use the same Lyapunov function when s becomes a little greater than s_0. It is clear that this results in a small variation of $\dot{V}(\xi, \eta)$, and hence this function may become positive only in a small neighbourhood of the origin. Therefore, the trajectories will intersect the closed level lines of the function V in the way

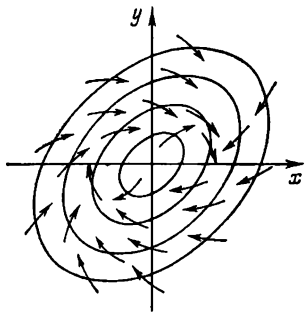


Fig. 169

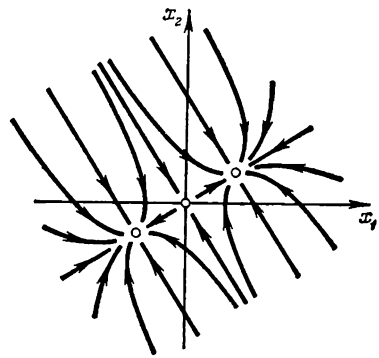


Fig. 170

shown in Fig. 169, that is only the level lines lying very close to the origin can be intersected by the trajectories from the inside to the outside. This means that although the zero solution becomes unstable every solution resulting from its small perturbation will remain in the vicinity of the origin. A typical phase diagram

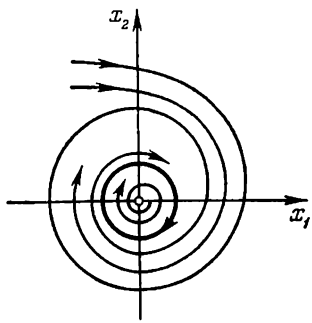


Fig. 171

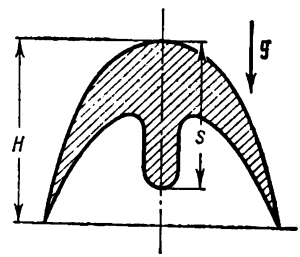


Fig. 172

after such a loss of stability is depicted in Fig. 170: the former stable nodal point has turned into a saddle point in whose vicinity two new stable nodes have appeared. In the similar case of a pair of pure imaginary roots of the characteristic equation on the boundary of the stability region a typical phase diagram appearing after the boundary has been crossed is shown in Fig. 171: the stable focal point has turned into an unstable one and has gene-

rated a small stable limit cycle to which the trajectories beginning far from the origin are spirally convergent.

In the examples studied above we dealt with a "safe" zone of the boundary of the stability region: a small deviation from it does not make the system go far away from the equilibrium state and only changes slightly its position and generates oscillations with small amplitude. For example, consider Fig. 172 where the axial section of a "cup with a leg" set forth from its bottom is depicted. The cup is shown as being turned upside-down and

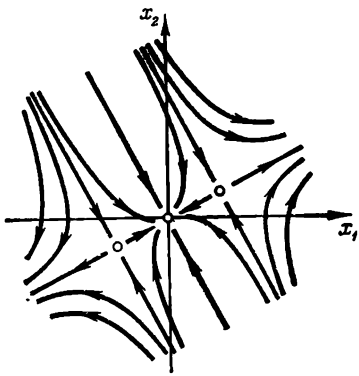


Fig. 173

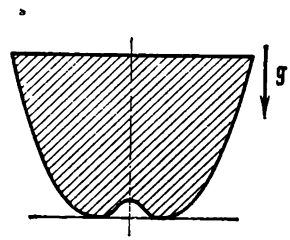


Fig. 174

supported by its brim; when the length s of the leg is less than the height H of the cup this equilibrium position is stable. If s becomes a little greater than H the position in which the axis of the cup is vertical becomes unstable but there appears a new ("oblique") stable state (determined to within a rotation) which does not differ considerably from the former state corresponding to $s < H$. But if the length of the leg becomes so large that the corresponding point s goes far away from the safe zone of the boundary of the stability region the system is no longer stable.

The case when the zero solution is unstable for $s = s_0$ is treated analogously. The ultimate result is that for the values of s which are a little smaller than s_0 the zero solution is (formally) stable but the "stability factor" is very small and therefore a substantial perturbation of the initial conditions may lead the system far away from the origin. A typical phase diagram corresponding to such a "weakly stable" solution is shown in Fig. 173 for the case on one zero root of the characteristic equation appearing when $s = s_0$. In the case of two pure imaginary roots we obtain a diagram similar to that in Fig. 171 with the reversed directions of the arrows. In these cases we deal with a "dangerous" zone of the boundary of the stability region. An example illustrating a stable equilibrium state in the vicinity of a dangerous zone of the boundary of the stability region is shown in Fig. 174. Generally spea-

king, as s decreases ($s < s_0$), the stability factor increases, that is the stability of the system "is improved".

In concrete problems, to investigate the character of stability of a system on the boundary of its stability region and to determine in this way safe and dangerous zones of the boundary we can carry out a computational experiment similar to the one described in Sec. 4.

For periodic systems we shall confine ourselves to studying critical cases for the simple scalar equation:

$$\dot{x} = a_k(t)\, x^k + a_{k+1}(t)\, x^{k+1} + \ldots \qquad (k = 2,\ 3,\ 4,\ \ldots) \qquad (32)$$

where all the coefficients $a_j(t)$ are periodic with period T. The situation is quite clear if

$$I = \int_0^T a_k(t)\, dt \neq 0$$

We then drop in equation (32) the terms of higher order of smallness after which the integration shows that

$$\frac{1}{[x(0)]^{k-1}} - \frac{1}{[x(T)]^{k-1}} \approx (k-1)\, I$$

that is

$$\left[\frac{x(T)}{x(0)}\right]^{k-1} \approx \{1 - (k-1)\, I\, [x(0)]^{k-1}\}^{-1}$$

It follows that for even k we always have instability while for odd k we have instability when $I > 0$ and stability when $I < 0$. If $I = 0$ the influence of the higher-order terms should be taken into account. To investigate this case we can use a substitution $y = x + \psi_k(t)\, x^k + \psi_{k+1}(t)\, x^{k+1} + \ldots$ with periodic coefficients to reduce equation (32) to the form $\dot{y} = b_k y^k + b_{k+1} y^{k+1} + \ldots$ with constant coefficients $b_k,\ b_{k+1},\ \ldots$. (In this procedure which we leave to the reader the coefficients $b_k,\ \psi_k(t),\ b_{k+1},\ \psi_{k+1}(t), \ldots$ are found in succession.) After this transformation the stability of the zero solution is determined by the first nonzero term on the right-hand side.

The problem of investigating a rest point in the plane for the case of two pure imaginary roots of the characteristic equation (see Sec. 2.4) reduces to studying the stability of the zero solution of an equation of form (32). Indeed, as was shown, the system in question is reduced, after an affine transformation of the plane has been made, to the form

$$\dot{x}' = \lambda y' + M(x',\ y'),\ \dot{y}' = -\lambda x' + N(x',\ y')$$

where the expansions of the functions M and N contain only

terms of order higher than the first. Passing to polar coordinates ρ, φ in the $x'y'$-plane we obtain

$$\left.\begin{aligned}\frac{d\rho}{dt} &= M(\rho\cos\varphi,\ \rho\sin\varphi)\cos\varphi + N(\rho\cos\varphi,\ \rho\sin\varphi)\sin\varphi\\ \frac{d\varphi}{dt} &= -\lambda + \frac{1}{\rho}[N(\rho\cos\varphi,\ \rho\sin\varphi)\cos\varphi - M(\rho\cos\varphi,\ \rho\sin\varphi)\sin\varphi]\end{aligned}\right\}$$

Dividing the first equation by the second, we finally receive an equation of type (32) with respect to $\rho(\varphi)$. (Let the reader analyse the implications of this fact.)

For more detail on critical cases see [71].

6. Special Classes of Mechanical Systems. Let us consider a potential autonomous system, with finite number of degrees of freedom and generalized coordinates $q_1, q_2, \ldots, q_n$, in the vicinity of an equilibrium position which is supposed to be at the origin. The kinetic energy $T(\dot{q}, q)$ of the system is a positive definite quadratic form in $\dot{q}_1, \dot{q}_2, \ldots, \dot{q}_n$, and the potential energy $U(q)$ has $q=0$ as its stationary point. The canonical equations of motion (see § VI.3) are written in the form

$$\frac{dp_j}{dt} = -\frac{\partial H}{\partial q_j},\quad \frac{dq_j}{dt} = \frac{\partial H}{\partial p_j} \qquad (j=1, 2, \ldots, n)$$

where

$$H(p, q) = T + U,\quad p_j = \frac{\partial(T-U)}{\partial \dot{q}_j} = \frac{\partial T}{\partial \dot{q}_j},\quad p=(p_1, \ldots, p_n),$$
$$q=(q_1, \ldots, q_n) \tag{33}$$

For a system of this type the law of conservation of energy holds: $H=\text{const}$ along every solution of the system.

These properties imply **Lagrange's theorem** (on stability): *if $U(q)$ has an isolated* (i.e. strict) *minimum for $q=0$ the equilibrium state $q=p=0$ is stable.* Indeed, here we can take $H(p, q)$ as a Lyapunov function and apply Lyapunov's theorem on stability (Sec. 3). If this minimum of the function $U(q)$ is nonisolated and is achieved on a manifold (M) it can easily be shown that the state of equilibrium is stable only with respect to small pulses while arbitrarily small perturbations of the initial data may result in a finite displacement over (M).

Let us write the expressions for the kinetic and potential energies in the vicinity of the equilibrium state $q=0$ in the form

$$T(\dot{q}, q) \equiv \frac{1}{2}\sum_{j,\,k} A_{jk}(q)\dot{q}_j\dot{q}_k = \frac{1}{2}\sum_{j,\,k} a_{jk}\dot{q}_j\dot{q}_k + o(1)|\dot{\mathbf{q}}|^2$$
$$(a_{jk} = A_{jk}(0))$$
$$U(q) = \frac{1}{2}\sum_{j,\,k} b_{jk}q_jq_k + o(|\mathbf{q}|^2)\quad (b_{jk}=\text{const})$$

where the quadratic forms on the right-hand sides are the principal terms. A linear change of variables $\mathbf{q} = \mathbf{H}\eta$ $(\mathbf{H} = (h_{jk}))$ yields $\dot{\mathbf{q}} = \mathbf{H}\dot{\eta}$, and hence both quadratic forms $\frac{1}{2}\sum_{j,\,k} a_{jk}\dot{q}_j\dot{q}_k$ and $\frac{1}{2}\sum_{j,\,k} b_{jk}q_jq_k$ are transformed with the aid of the same matrix $\mathbf{H}$. The first form being positive definite, it is always possible (see Sec. IV.2.4) to choose the transformation matrix $\mathbf{H}$ in such a way that in the new coordinates the principal terms of the energies take the form

$$T = \frac{1}{2}\sum_j \dot{\eta}_j^2, \quad U = \frac{1}{2}\sum_j \mu_j\eta_j^2$$

where the structure of the last quadratic form is the same as that of the form $(\mathbf{Bq}, \mathbf{q})$ $(\mathbf{B} = (b_{jk}))$ and μ_j $(j = 1, 2, \ldots, n)$ are the eigenvalues of the matrix $\mathbf{A}^{-1}\mathbf{B}$ $(\mathbf{A} = (a_{jk}))$. These new coordinates are called **normal** or **principal coordinates** of the system.

In the normal coordinates the linearized equations of motion corresponding to system (33) take the simple form

$$\frac{d\xi_j}{dt} = -\mu_j\eta_j, \quad \frac{d\eta_j}{dt} = \xi_j \qquad (j = 1, 2, \ldots, n) \tag{34}$$

(in this case the generalized momentum ξ_j coincides with $\dot{\eta}_j$ to within infinitesimals of higher order). To write down explicitly the characteristic determinant it is convenient to arrange the unknown functions in the sequence $\xi_1, \eta_1, \xi_2, \eta_2, \ldots, \xi_n, \eta_n$. The roots of the characteristic equation turn out to be equal to $\pm\sqrt{-\mu_1}, \pm\sqrt{-\mu_2}, \ldots, \pm\sqrt{-\mu_n}$ (check it up!). Now, applying the conclusion drawn in the second paragraph of this section and the results of Sec. 4 we can assert that *if* $\mu_j > 0$ $(j = 1, 2, \ldots, n)$ *the equilibrium state under consideration is stable; if at least one of the values* μ_j *is nonpositive this state is unstable.*

(It can be proved that, generally, if $U(q)$ does not have a minimum for $q = 0$, even a nonstrict one, the equilibrium solution $q = 0$ is unstable). Since the positivity of all μ's is equivalent to the condition that the matrix $\mathbf{A}^{-1}\mathbf{B}$ is positive definite, this property can be tested by means of the Sylvester theorem (see Sec. IV.2.3).

If there is a value $\mu_l > 0$ the particular solution

$$\eta_l = M\sin(\sqrt{\mu_l}t + \varphi), \quad \eta_j \equiv 0 \quad (j \neq 0)$$

of system (34) determines a **normal mode of vibration** or, simply, a **normal mode** (also spoken of as a **principal oscillation**) of the linearized system in question. In the original coordinates this normal mode is described by the formulas

$$q_j = Mh_{jl}\sin(\sqrt{\mu_l}t + \varphi) \qquad (j = 1, 2, \ldots, n)$$

and thus it is a **monofrequent vibration** (a **natural mode of vibration**) with **natural frequency** $\sqrt{\mu_l}$. (If several natural frequencies coincide the choice of the corresponding normal coordinates is not uniquely determined, and after such a choice has been made in a certain manner the class of natural modes of vibration is (formally) wider than the class of normal modes (why?).) Substituting these expressions of the solutions into the linearized equations of motion which can be written as

$$\sum_k (a_{jk}\ddot{q}_k + b_{jk}q_k) = 0 \qquad (j = 1, 2, \ldots, n)$$

or, equivalently, substituting the expression $\mathbf{q} = M\mathbf{h}_l \sin(\sqrt{\mu_l}t + \varphi)$, where $\mathbf{h}_l = \begin{pmatrix} h_{1l} \\ \cdot \\ \cdot \\ \cdot \\ h_{nl} \end{pmatrix}$, into the equation

$$\mathbf{a}\ddot{\mathbf{q}} + \mathbf{b}\mathbf{q} = \mathbf{0}$$

($\mathbf{a} = (a_{jk}) = \mathbf{A}$, $\mathbf{b} = (b_{jk}) = \mathbf{B}$) we see that the vector $\mathbf{h}_l$ must satisfy the relation

$$(\mu_l \mathbf{a} - \mathbf{b})\,\mathbf{h}_l = \mathbf{0}, \text{ that is } (\mathbf{a}^{-1}\mathbf{b})\,\mathbf{h}_l = \mu_l \mathbf{h}_l$$

The definition of the normal coordinates implies that

$$\mathbf{H}^*\mathbf{a}\mathbf{H} = \mathbf{I}, \quad \mathbf{H}^*\mathbf{b}\mathbf{H} = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_n)$$

whence it follows (why?) that

$$(\mathbf{a}\mathbf{h}_j, \mathbf{h}_k) = \delta_{jk}, \quad (\mathbf{b}\mathbf{h}_j, \mathbf{h}_k) = 0 \ (j \neq k), \quad (\mathbf{b}\mathbf{h}_j, \mathbf{h}_j) = \mu_j \qquad (35)$$

where $\mathbf{h}_s$ $(s = 1, \ldots, n)$ is the sth column of the matrix $\mathbf{H}$. (By the way, the scheme given in Sec. IV.1.7 can be applied to construct, on the basis of the above results, the extremal theory for natural frequencies; for instance, it can be proved that μ_1, which is the greatest of the frequencies μ_j, is equal to $\max_{\mathbf{h}} \frac{(\mathbf{bh}, \mathbf{h})}{(\mathbf{ah}, \mathbf{h})}$, the subsequent frequency being obtained by adding the condition $(\mathbf{ah}, \mathbf{h}_1) = 0$, etc.) The general form of vibrations in the system in question is expressed as the sum

$$\mathbf{q} = \sum_{j=1}^{n} \mathbf{q}_j(t), \quad \mathbf{q}_j(t) = M_j \mathbf{h}_j \sin(\sqrt{\mu_j}\, t + \varphi_j)$$

of normal modes where all M_j and φ_j are arbitrary constants. By (35), it follows that

$$\tfrac{1}{2}(\mathbf{a}\dot{\mathbf{q}}, \dot{\mathbf{q}}) = \tfrac{1}{2}\sum_j (\mathbf{a}\dot{\mathbf{q}}_j, \dot{\mathbf{q}}_j) = \tfrac{1}{2}\sum_j \mu_j M_j^2 \cos^2(\sqrt{\mu_j}\, t + \varphi_j)$$

and

$$\frac{1}{2}(\mathbf{bq},\ \mathbf{q}) \equiv \frac{1}{2}\sum_j (\mathbf{bq}_j,\ \mathbf{q}_j) = \frac{1}{2}\sum_j \mu_j M_j^2 \sin^2(\sqrt{\mu_j}\, t + \varphi_j)$$

Consequently, at each moment the kinetic and potential energies of vibrations are the sums of the corresponding energies of the normal modes.

The number $\varkappa$ of the negative quantities μ_j characterizes the degree of instability of the equilibrium state $q=0$. (Fig. 175 illustrates the cases of different degrees of instability for a material point moving without friction in a surface under the action of gravitation.) In the case $\varkappa = n$ we have absolute (complete) instability.

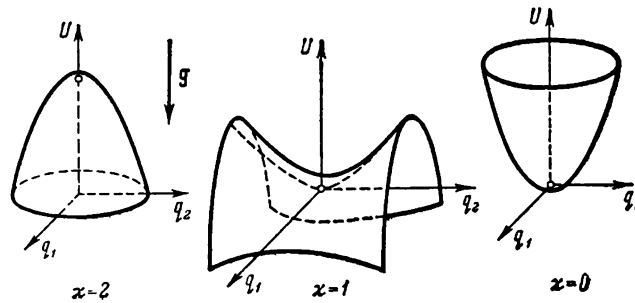


Fig. 175

It appears evident that every additional constraint imposed on the coordinates of the equilibrium position $q=0$ may either leave the number $\varkappa$ unchanged or reduce it by unity. (This is clearly seen in the middle Fig. 175 if we draw different planes through the U-axis and require that the moving point should remain in them.) Consequently, if $\varkappa < n$ and $\mu_j \neq 0$, $j = 1, 2, \ldots, n$, the introduction of $\varkappa$ constraints may result in a stable system. A completely unstable system remains unstable irrespective of the constraints imposed on it.

The above assertion on the introduction of additional constraints can be proved rigorously in the following way. First of all, according to (34), we have $\ddot{\xi}_j = -\mu_j \xi_j$, and therefore the numbers μ_j can be determined if we find a nontrivial solution of the linearized system (in any coordinates not necessarily in normal ones) in the form $\eta = e^{\omega t}\mathbf{c}$ and then put $\mu = -\omega^2$. Now suppose that there is a constraint of the form $\sum_j A_j \eta_j = 0$. We then arrive at the problem of determining a stationary value of the integral

$\int (T-U)\,dt$ under one finite constraint. Such a problem (see Sec. VI.1.12) leads to the system of Euler's equations

$$\ddot{\eta}_j+\mu_j\eta_j+\lambda(t)\,A_j=0 \qquad (j=1, 2, \ldots, n)$$

The substitution of a solution of the form $\eta_j=c_je^{\omega t}$ into these equations and into the equation of the constraint results in the relation (check it up!)

$$\sum_j \frac{A_j^2}{\mu-\mu_j}=0 \quad (\mu=-\omega^2) \tag{36}$$

For simplicity, let all $A_j\neq 0$ and all μ_j be distinct (the special cases for which these conditions are violated can be investigated by means of the passage to the limit). Then the graph of the left

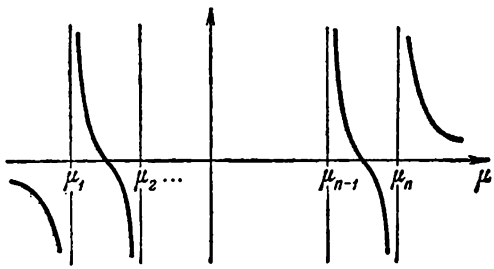


Fig. 176

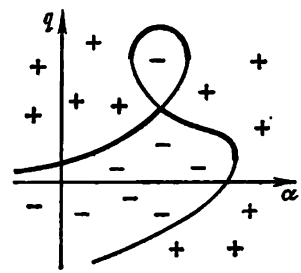


Fig. 177

member of (36) is of the form shown in Fig. 176 (why?). Consequently, all $n-1$ roots of equations (36) are real, distinct and alternate with the values of μ_j, whence follows the desired proposition.

If the system under investigation depends on parameters $\alpha_1, \ldots, \alpha_k$ we encounter the problem of determining regions in the space of the parameters to which stable equilibrium states correspond. If the potential energy of the system is expressed by a function $U(q_1, \ldots, q_n, \alpha_1, \ldots, \alpha_k)$ the equilibrium positions are specified by the equations

$$U_{q_1}'=0, \quad U_{q_2}'=0, \quad \ldots, \quad U_{q_n}'=0 \tag{37}$$

whose solution, generally speaking, determines a collection of k-dimensional surfaces in the $(n+k)$-dimensional space of the variables q, α. With each value $\alpha=(\alpha_1, \ldots, \alpha_k)$ a number of values of $q=(q_1, \ldots, q_n)$ (possibly, none) satisfying equations (37) are associated, the number $\varkappa$ which characterizes the degree of instability of the corresponding stationary state being equal to the number of negative roots of the equation

$$\det(U''_{q_iq_j}-\lambda\delta_{ij})=0 \tag{38}$$

(why?). If all the functions involved and their derivatives are continuous the number $\varkappa$ can only change when the condition

$$\det\left(U''_{q_iq_j}\right)=0$$

holds. But the points at which this relation is fulfilled are those at which the sufficient conditions for the existence of the implicit function $q(\alpha)$ are violated. Therefore, when investigating the stability we should take into account not only the bifurcation conditions for this function but also the variation of the signs of the roots of equation (38). However, in some practically important cases it is sufficient to determine the sign of the determinant $\Delta=$ $=\det\left(U''_{q_iq_j}\right)$; for instance, in the region where $\Delta<0$ the system is sure to be unstable (why?).

The simplest case occurs when $n=k=1$. Then, instead of (37), we deal with the equation $U'_q(q,\ \alpha)=0$ which determines a line of equilibrium in the αq-plane, and it suffices to determine the sign of U''_{qq} on this line to judge upon the stability. (For instance, let the reader verify that if the sign of the derivative U'_q changes as is shown in Fig. 177, the stable parts of the line of equilibrium are those shown in heavy line in the figure.)

If $k=1$ a simple situation similar to the above sometimes takes place even when $n>1$. For example, let it be possible to express $q_2, \ldots, q_n$ in terms of q_1 and α by solving the last $n-1$ equations (37). Substituting the expressions thus found into U'_{q_1} we obtain a function of q_1 and α which will be denoted by $f(q_1,\ \alpha)$. Then in the αq_1-plane we have the equilibrium line $f=0$. It can be verified by direct computation that

$$\frac{\partial f}{\partial q_1}=\det\left(U''_{q_iq_j}\right)D^{-1}$$

where D is the determinant of the $(n-1)$th order obtained from $\det\left(U''_{q_iq_j}\right)$ by deleting the first row and the first column. This enables us to find the parts of the equilibrium line on which the system is sure to be unstable (because of the condition $\Delta<0$).

Now we proceed to study the case when nonpotential perturbation forces act upon a given canonical system. Let us rewrite the unperturbed system of canonical equations (33) in the form

$$\frac{dp_j}{dt}=-\frac{\partial T}{\partial q_j}-\frac{\partial U}{\partial q_j},\quad \frac{dq_j}{dt}=\frac{\partial T}{\partial p_j}\qquad (j=1,\ 2,\ \ldots,\ n)$$

This indicates that the nonpotential forces we have spoken of should be added to the right member of the first equations. Confining ourselves to normal coordinates and linearized equations we

shall begin with the so-called **dissipative forces** of the form

$$Q_j = \sum_{k=1}^{n} c_{jk}\xi_k \qquad (c_{jk} = c_{kj})$$

where the quadratic form $\sum_{j,k} c_{jk}\xi_j\xi_k$ is positive semidefinite or positive definite. In the latter case we speak of **total dissipation.** It is readily seen that for such forces we have

$$\frac{d}{dt}(U+T) = -\sum_{j,k} c_{jk}\xi_j\xi_k \leqslant 0$$

Therefore, *if the state of equilibrium* $\eta = 0$ *is stable for the original system* (that is the function $U(\eta)$ has a minimum for $\eta = 0$) *it is also stable for the system with additional dissipative forces* (this is implied by the stability theorem given in Sec. 3). *In the case of total dissipation the solution* $\eta = 0$ *is asymptotically stable* (this follows from the theorem on asymptotic stability of Sec. 3 if we take into account the comment about the arcs of trajectories). It can also be proved that the application of dissipative forces does not change the character of isolated nonstable stationary points and of nonisolated unstable equilibrium states for which $\varkappa > 0$.

The so-called **gyroscopic forces**

$$Q_j = \sum_{k=1}^{n} g_{jk}\xi_k \text{ where } g_{jk} = -g_{kj} \qquad (j,\ k = 1, 2, \ldots, n)$$

form another important class of forces. A feature of such a force is that it performs zero work for any displacement of the system:

$$dA = \mathbf{Q}\cdot d\eta = \sum_j Q_j\xi_j\,dt = \sum_{j,k} g_{jk}\xi_j\xi_k\,dt = 0 \qquad \left(\mathbf{Q} = \begin{pmatrix} Q_1 \\ \vdots \\ Q_n \end{pmatrix}\right)$$

If gyroscopic and dissipative forces are applied simultaneously we have

$$\frac{d}{dt}(U+T) = -\sum_{j,k} c_{jk}\xi_j\xi_k$$

Hence, as above, we conclude that every stable state of equilibrium of the original system remains stable under the action of such forces, and in the case of total dissipation it becomes asymptotically stable. It can be proved that in the case of total dissipation an unstable nonisolated stationary state remains unstable when gyroscopic and dissipative forces are applied to the system simultaneously.

If the additional forces are purely gyroscopic the instability is retained if the number $\varkappa$ is odd while for even $\varkappa$ it may turn into nonasymptotic stability. For instance, take the case $n = 2$. Then

$$U = -\frac{\alpha}{2}(\eta_1^2 + \eta_2^2) \quad (\alpha > 0), \quad Q_1 = g\xi_2, \quad Q_2 = -g\xi_1$$

and the dependence of the solutions on time t is due to the factor $e^{\lambda t}$ where the exponent λ satisfies the equation

$$\lambda^4 + (g^2 - 2\alpha)\lambda^2 + \alpha^2 = 0$$

(check it up!). All the roots of this equation are pure imaginary for $g > 2\sqrt{\alpha}$, which implies the stability. A sufficiently large gyroscopic force (directed perpendicularly to the trajectory of motion) generates quasi-periodic oscillations of the system about the equilibrium position. However, the application of arbitrarily small constantly operating dissipative forces destroys, in a sufficiently long time of motion, these stable oscillations. (Let the reader analyse the realization of all these considerations by considering the example of a spinning top.)

In engineering a complicated mechanical system is usually composed of a number of "partial" systems by means of rigid or elastic connections. As an example of this kind we can take two pendulums connected by a rod (a rigid coupling) or by a spring (an elastic coupling).

Suppose that two autonomous conservative linear systems with generalized coordinates $q' = (q'_1, q'_2, \ldots, q'_{n'})$ and $q'' = (q''_1, q''_2, \ldots, q''_{n''})$ whose kinetic and potential energies are, respectively, $(\mathbf{a'\dot{q}'}, \dot{\mathbf{q}}')$, $(\mathbf{b'q'}, \mathbf{q}')$ and $(\mathbf{a''\dot{q}''}, \dot{\mathbf{q}}'')$, $(\mathbf{b''q''}, \mathbf{q}'')$ are connected with linear elastic couplings. This results in a system with $n' + n''$ degrees of freedom and generalized coordinates $q = (q', q'')$. If the new couplings are inertialess (i.e. their elements have negligibly small masses) the kinetic and potential energies of the composite system can be written in the form $(\mathbf{a\dot{q}}, \dot{\mathbf{q}})$ and $(\mathbf{bq}, \mathbf{q})$ where

$$\mathbf{a} = \begin{pmatrix} \mathbf{a'} & \mathbf{0} \\ \mathbf{0} & \mathbf{a''} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \mathbf{b'} & \mathbf{r} \\ \mathbf{r^*} & \mathbf{b''} \end{pmatrix}$$

and $\mathbf{r}$ is an $(n' \times n'')$ **coupling matrix** for the partial systems. Besides, the inequality

$$(\mathbf{bq}, \mathbf{q}) \geqslant (\mathbf{b'q'}, \mathbf{q'}) + (\mathbf{b''q''}, \mathbf{q''})$$

must hold (why?), which is equivalent to the condition

$$(\mathbf{rq''}, \mathbf{q'}) \geqslant 0$$

It follows that if the constituent systems are stable the composite system is stable as well. It is possible to prove that the natural

frequencies of the composite system lie between the least and the greatest partial frequencies, that is natural frequencies of the constituent systems. If the connecting forces are small (which means that the elements of the matrix $\mathbf{r}$ are relatively small) we can use the perturbation method (Sec. IV.4.10) to determine the natural frequencies of the composite system by taking the partial frequencies as an initial approximation.

7. Automatic Control Systems. In recent years much emphasis has been laid upon studying the so-called *automatic control systems*; they are described by nonlinear equations which retain some features of linear systems, which makes it possible to obtain far-reaching results in their investigation. The Soviet scientist A. I. Lurye (born in 1901) was the first to study equations of this type. At present there are a number of books devoted to this problem (e.g. see [1], [83], [84] and [89]).

Suppose that there is a controlled plant which is linear and is described by a system of equations $\dot{\mathbf{x}} = \mathbf{Ax}$, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ (the justification of this assumption lies in the fact that usually in the operation process the deviation of the object from its equilibrium position must not be large). The regulation is performed by means of a **control action (steering signal** or **controlling influence)** ξ which, for the sake of simplicity, is supposed to be scalar. Thus, the **controlled object** is described by the equation $\dot{\mathbf{x}} = \mathbf{Ax} - \xi\mathbf{b}$ where $\mathbf{b}$ is a given vector. The controlling influence is generated by a **feedback signal** σ (which is assumed to be scalar) according to the formula $\xi = \varphi(\sigma)$ (in the case of the so-called **direct control**) or $\dot{\xi} = \varphi(\sigma)$ (in the case of the so-called **indirect control**). Here $\varphi(\sigma)$ is a given **controller characteristic**. The feedback signal is specified by the state of the controlled object and by the controlling influence: $\sigma = \mathbf{c}^*\mathbf{x} - \rho\xi$ where $\mathbf{c}$ is a given vector and ρ a given scalar.

For definiteness, let us consider the case of indirect control. The complete system of equations is then written in the form

$$\dot{\mathbf{x}} = \mathbf{Ax} - \xi\mathbf{b}, \quad \dot{\xi} = \varphi(\sigma), \quad \sigma = \mathbf{c}^*\mathbf{x} - \rho\xi \tag{39}$$

and involves only one nonlinear term $\varphi(\sigma)$. In real systems the controller may work in an essentially nonlinear regime; for instance, its characteristic may be of the form $\varphi(\sigma) = k \operatorname{sgn} \sigma$ $(k > 0)$, and therefore the general theory naturally includes such nonlinear terms. In what follows, without further stipulations, we shall suppose that the characteristic chosen possesses the following properties:

$$\operatorname{sgn} \varphi(\sigma) = \operatorname{sgn} \sigma, \quad \int_0^{\infty} \varphi(\sigma)\, d\sigma = \infty, \quad \int_{-\infty}^{0} \varphi(\sigma)\, d\sigma = -\infty$$

One of the most thoroughly investigated questions in the theory of automatic control systems is the problem posed by A. I. Lurye:

to find out what are the matrix $\mathbf{A}$ and the coefficients $\mathbf{b}$, $\mathbf{c}$ and ρ for which the zero solution of system (39) possesses global asymptotic stability (Sec. 3) for any choice of the characteristic $\varphi(\sigma)$. If $\mathbf{A}$, $\mathbf{b}$, $\mathbf{c}$ and ρ are such that this condition is fulfilled system (39) is said to be **absolutely stable**. Since in real devices the control signal can be arbitrarily small it is natural to suppose that the original system $\dot{\mathbf{x}}=\mathbf{Ax}$ is asymptotically stable as well.

To investigate this problem we shall pass from the sought-for functions $\mathbf{x}$ and ξ to the new unknown functions $\mathbf{y}$ and σ connected with the former by the formulas

$$\mathbf{y}=\mathbf{Ax}-\xi\mathbf{b}, \quad \sigma=\mathbf{c}^*\mathbf{x}-\rho\xi \tag{40}$$

Then the system of differential equations takes the form (check it up!)

$$\dot{\mathbf{y}}=\mathbf{Ay}-\varphi(\sigma)\,\mathbf{b}, \quad \dot{\sigma}=\mathbf{c}^*\mathbf{y}-\rho\varphi(\sigma) \tag{41}$$

The transformation must of course be nondegenerate, that is

$$\begin{vmatrix} \mathbf{A} & -\mathbf{b} \\ \mathbf{c}^* & -\rho \end{vmatrix} \neq 0 \tag{42}$$

Now we need the following general formula for determinants which we give without proof: if $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are, respectively, some $(n\times n)$, $(n\times m)$, $(m\times n)$ and $(m\times m)$ matrices and $\det \mathbf{A}\neq 0$ then

$$\det \begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix} = \det \mathbf{A}\cdot\det(\mathbf{D}-\mathbf{CA}^{-1}\mathbf{B}) \tag{43}$$

This indicates that inequality (42) can be rewritten as the condition

$$\rho \neq \mathbf{c}^*\mathbf{A}^{-1}\mathbf{b} \tag{44}$$

In what follows, relation (44) will be supposed to be fulfilled. It implies, in particular, that system (41) possesses a single equilibrium position, namely $\mathbf{y}=\mathbf{0}$, $\sigma=0$ (why?).

To derive sufficient conditions for absolute stability of system (41) we shall take the Lyapunov function of the form

$$V(\mathbf{y}, \sigma)=\mathbf{y}^*\mathbf{By}+\int_0^{\sigma}\varphi(\sigma_1)\,d\sigma_1 \tag{45}$$

where $\mathbf{B}$ is a positive definite symmetric matrix to be specified later. $V(y, \sigma)$ is a continuous positive definite function for which $V(\infty)=\infty$. The direct computation shows that (check it up!)

$$-\dot{V}=\mathbf{y}^*\mathbf{Cy}+\rho\varphi^2(\sigma)+2\varphi(\sigma)\,\mathbf{d}^*\mathbf{y} \tag{46}$$

where

$$\mathbf{C}=-(\mathbf{A}^*\mathbf{B}+\mathbf{BA}), \quad \mathbf{d}=\mathbf{Bb}-\frac{1}{2}\mathbf{c} \tag{47}$$

For the theorem on asymptotic stability in the large (given in Sec. 3) to be applicable to (41) and function (45) it is necessary that the right member of (46), which is a quadratic form in $y_1, \ldots, y_n, \varphi(\sigma)$ (here lies the resemblance between the nonlinear system in question and linear systems!), should be positive definite. The matrix of this form being $\begin{pmatrix} \mathbf{C} & \mathbf{d} \\ \mathbf{d}^* & \rho \end{pmatrix}$, the Sylvester theorem and formula (43) indicate that this form is positive definite if and only if the matrix $\mathbf{C}$ is positive definite and the inequality

$$\rho > \mathbf{d}^*\mathbf{C}^{-1}\mathbf{d} = \left(\mathbf{b}^*\mathbf{B} - \frac{1}{2}\mathbf{c}^*\right)\mathbf{C}^{-1}\left(\mathbf{B}\mathbf{b} - \frac{1}{2}\mathbf{c}\right) \tag{48}$$

is fulfilled.

As was shown in Sec. 4, for any symmetric positive definite matrix $\mathbf{C}$ there exists exactly one symmetric matrix $\mathbf{B}$ satisfying the first equality (47); besides, the matrix $\mathbf{B}$ is positive definite as well and therefore can be regarded as a function of the form $\mathbf{B}=\mathbf{B}(\mathbf{C})$. Thus, *the conditions of the theorem on global asymptotic stability are fulfilled for the Lyapunov function of the given form* (45) *if and only if there is a positive definite symmetric matrix* $\mathbf{C}$ *satisfying inequality* (48), and consequently *the existence of* $\mathbf{C}$ *is sufficient for system* (39) *to be absolutely stable.*

It can be proved that for any asymptotically stable matrix $\mathbf{A}$ (that is one whose all eigenvalues have negative real parts) and every symmetric positive definite matrix $\mathbf{C}$ the vector $\mathbf{d}$ constructed above satisfies inequality $\mathbf{d}^*\mathbf{C}^{-1}\mathbf{d} > \mathbf{c}^*\mathbf{A}^{-1}\mathbf{b}$, and therefore the verification of condition (44) can be simplified. Consider the particular case when $\mathbf{A} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with all λ_j's real and negative. Let us choose the matrix $\mathbf{C}$ in the special form $\mathbf{C} = \operatorname{diag}(\alpha_1, \ldots, \alpha_n)$ where all α_j's are positive. Then

$$\mathbf{B} = -\frac{1}{2}\operatorname{diag}\left(\frac{\alpha_1}{\lambda_1}, \ldots, \frac{\alpha_n}{\lambda_n}\right), \quad d_j = -\frac{1}{2}\left(\frac{\alpha_j}{\lambda_j}b_j + c_j\right),$$

$$\mathbf{d}^*\mathbf{C}^{-1}\mathbf{d} = \frac{1}{4}\sum_j \frac{1}{\alpha_j}\left(\frac{\alpha_j}{\lambda_j}b_j + c_j\right)^2$$

Let us choose the coefficients α_j in such a way that the right-hand side of the last equality becomes as small as possible. It can be shown (let the reader prove it) that the least value is equal to $\sum_j{}' \frac{b_j c_j}{\lambda_j}$ where the summation extends only over the positive summands. Thus, by (48), we now conclude that the condition $\rho > \sum_j{}' \frac{b_j c_j}{\lambda_j}$ is sufficient for the system in question to be absolutely stable.

In 1961 the Rumanian mathematician V. M. Popov applied Fourier's transformation to obtain a stronger condition for absolute

stability which can be tested in a simpler way. Here we present, without proof, this result (in connection with system (41)) under the additional requirement $\rho \neq \mathbf{c}^*\mathbf{b}$. The Lyapunov function used by Popov has a more general form than (45); it is equal to the sum of the expression $\beta \int_0^{\sigma} \varphi(\sigma_1)\, d\sigma_1$ and a quadratic form with respect to the variables $y_1, \ldots, y_n, \sigma$. It was proved that it is possible to choose such a function satisfying the conditions of the theorem on asymptotic stability in the large if and only if there exists a number $q \geqslant 0$ for which

$$\operatorname{Re}\left[(1+i\omega q)\,\mathbf{c}^*\mathbf{A}^{-1}(i\omega\mathbf{I}-\mathbf{A})^{-1}\mathbf{b}\right]+q(\rho-\mathbf{c}^*\mathbf{b})\geqslant 0 \qquad (49)$$

for all real ω. This condition, like (48), is only sufficient for the absolute stability, but since the Lyapunov function taken here belongs to a wider class the new condition is "closer" to a necessary one (why?). Besides, condition (49) admits of a simple geometric interpretation. Consider the line (l) in the x y-plane with parametric equations

$$x=\omega \operatorname{Im}\left[\mathbf{c}^*\mathbf{A}^{-1}(i\omega\mathbf{I}-\mathbf{A})^{-1}\mathbf{b}\right]-\rho+\mathbf{c}^*\mathbf{b}$$
$$y=\operatorname{Re}\left[\mathbf{c}^*\mathbf{A}^{-1}(i\omega\mathbf{I}-\mathbf{A})^{-1}\mathbf{b}\right]$$

ω being the parameter ($-\infty < \omega < \infty$). Then condition (49) reads: there exists $q \geqslant 0$ such that $y(\omega) - qx(\omega) \geqslant 0$. Geometrically, this is equivalent to the requirement that the line (l) should entirely lie in the half-plane $y \geqslant qx$.

For unstable controlled objects we shall limit ourselves to indicating that *it is possible to stabilize such an object with the help of a vector controlling influence*. By analogy with the above passage from the system $\dot{\mathbf{x}} = \mathbf{Ax}$ to system (41), this assertion can be reformulated as follows: *for any* $(n \times n)$ *matrix* $\mathbf{A}$ (all the quantities involved being considered real) *it is possible to choose, respectively,* $(n \times m)$, $(m \times n)$ and $(m \times m)$ *matrices* $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ (the number m is also chosen in an appropriate manner) *such that all the eigenvalues of the block matrix*

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \qquad (50)$$

have negative real parts. Such an operation on $\mathbf{A}$ resulting in (50) is spoken of as **bordering** the matrix $\mathbf{A}$. Consequently, *any matrix can be made asymptotically stable by means of bordering*.

For the sake of simplicity, let us assume that all the Jordan submatrices of the matrix $\mathbf{A}$ are of the first order (in the general case the construction is analogous). Then, by virtue of Sec. IV.3.4, it is possible to choose a matrix $\mathbf{H}$ such that the matrix $\mathbf{A}' = \mathbf{H}^{-1}\mathbf{AH}$

has a quasi-diagonal form with first-order matrices of the form (λ_j) and second-order matrices of the form $\begin{pmatrix} \mu_j & \nu_j \\ -\nu_j & \mu_j \end{pmatrix}$ along its principal diagonal. If $\lambda_j \geqq 0$, the bordering of the matrix (λ_j) can be performed by passing to the matrix

$$\begin{pmatrix} \lambda_j & 1 \\ u & v \end{pmatrix},$$

where, as can easily be shown (prove it!), the parameters u and v can be selected in such a way that the coefficients of the characteristic equation (and therefore its roots) are equal to any given numbers. Similarly, for a second-order matrix of the above type with $\mu_j \geqq 0$ the bordering can be achieved in the following way:

$$\begin{pmatrix} \mu_j & \nu_j & 0 \\ -\nu_j & \mu_j & 1 \\ u & v & w_j \end{pmatrix}$$

(check it up!). After each of the Jordan submatrices has been bordered as indicated here we can pass to the corresponding asymptotically stable bordering of the matrix $\mathbf{A}'$ by interchanging simultaneously, in the appropriate manner, pairs of rows and columns with the same indices. This results in a bordered matrix $\mathbf{A}'$ whose all eigenvalues have negative real parts. Finally, to obtain the required bordering of the matrix $\mathbf{A}$ we use the formula

$$\begin{pmatrix} \mathbf{H} & 0 \\ 0 & \mathbf{I}_m \end{pmatrix} \begin{pmatrix} \mathbf{A}' & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{H} & 0 \\ 0 & \mathbf{I}_m \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A} & \mathbf{HB} \\ \mathbf{CH}^{-1} & \mathbf{D} \end{pmatrix}$$

(This proof incidentally shows that by bordering a given matrix we can always construct a matrix of a sufficiently high order whose eigenvalues are equal to any given numbers.) In practical applications it is of course advisable to make m as small as possible.

For various stabilization problems we refer the reader to [7], Chapter II.

In conclusion we mention here the problem posed by M. A. Aizerman which is in a certain sense close to the question of possibility to judge upon the stability of a system with variable coefficients dependent on t with the help of the corresponding system with constant coefficients involving, instead of t, a parameter τ (see Sec. 4). We shall take a system of differential equations containing only one nonlinear function of one independent variable, say of form (4.1). Let the nonlinear term be of the form $\varphi(\sigma)$ $\psi(\sigma)\sigma$ where $\alpha < \psi(\sigma) < \beta$. Besides, suppose that the substitution of any linear function of the form $\psi_0\sigma$ $(\alpha < \psi_0 < \beta)$ for the nonlinear function $\varphi(\sigma)$ results in an asymptotically stable, in the large,

linear system. The **problem of M. A. Aizerman** is to find out whether the original system with nonlinear term $\varphi(\sigma)$ is then also asymptotically stable in the large. It is clear that in the cases when the answer to the question is positive we obtain a simple test for stability.

This problem has been treated by many scientists (in particular, see [111]). It turned out that in the general case the answer is negative although there exist particular classes of systems for which it is positive.

For example, take the system

$$\dot{x}=\psi(x)x+ay,\ \dot{y}=bx+cy\ \ (\alpha\leqslant\psi(x)\leqslant\beta) \tag{51}$$

For the linear system obtained from (51) by substituting a fixed parameter $\psi_0\ (\alpha\leqslant\psi_0\leqslant\beta)$ for $\psi(x)$ the condition of asymptotic stability is written in the form

$$\psi_0+c<0,\ \psi_0 c-ab>0\ \ (\alpha\leqslant\psi_0\leqslant\beta) \tag{52}$$

(check it up!). Consider the positive definite Lyapunov function

$$V(x,\ y)=\int_0^x[\psi(x_1)c-ab]\,x_1dx_1+\frac{1}{2}(cx-ay)^2 \tag{53}$$

equal to ∞ at infinity. Its total time derivative is equal to

$$\dot{V}=[\psi(x)c-ab]\,x\,[\psi(x)x+ay]+$$
$$+(cx-ay)\{c[\psi(x)x+ay]-a(bx+cy)\}\equiv x^2[\psi(x)+c]\,[\psi(x)c-ab]$$

From conditions (52) (which are also fulfilled for the original variable coefficient $\psi(x)$) it follows that $\dot{V}\leqslant 0$ and that $\dot{V}=0$ only on the y-axis which does not contain any entire arc of a trajectory for $a\neq 0$. Thus, by virtue of Sec. 3, under the assumption $a\neq 0$ the asymptotic stability in the large has been proved for system (51). For $a=0$ the same result is obtained if, instead of (53), we take the function $V=Ax^2+y^2$ with a sufficiently large coefficient A (elaborate the proof!). Hence, for systems (51) the answer to the Aizerman problem is positive.

8. "Technical" Stability. Let us come back to the general system (3). Suppose that we are given an estimate or a system of estimates for the perturbations of the initial data. For instance, it may be indicated that $\max_j|x_j(t_0)|\leqslant h$ (one estimate) or $|x_j(t_0)|\leqslant h_j;\ j=1,\ 2,\ \ldots,\ n$ (this is a system of estimates). Besides, if the right member of (3) is also perturbed we can be given an estimate or a system of estimates for this perturbation on a given time interval $(t_0,\ T)$ where T can be finite or infinite. Finally, assume that there is an estimate or a system of estimates which must be satisfied by the perturbations of the solutions.

In these circumstances it is sometimes said that the unperturbed solution is "technically" stable with respect to all these conditions (estimates) if, given any perturbations of the initial data and of the right-hand sides of the equations satisfying the given estimates, the corresponding perturbations of the solutions satisfy the estimates imposed on them for the given time interval.

We think that the term "technical" stability and the setting of the problem connected with it are inappropriate in the general case since they rarely yield valuable practical results while a simpler computational experiment more often proves effective. (To justify this scheme it is sometimes said that an engineer is always interested in a finite interval of time and finite perturbations whereas the classical stability theory deals with infinite intervals and infinitesimal perturbations; such a comment is connected with misunderstanding, because the classical setting of the problem does not exclude the application to real time intervals and perturbations. E.g. see [107], Sec. XV.22 on these questions.)

But despite this general remark there are of course some cases when the above-mentioned guaranteed estimates for the perturbations of solutions prove useful. If only initial data are perturbed the problem reduces to the estimation of the solution of the given differential equation with the given estimate for the initial data. Such an estimation can be derived from the exact solution (provided it is known), from first integrals of the system (e.g. see [107], Sec. XV.13) and by numerical integration with randomly chosen initial values.

The method of Lyapunov's functions is also sometimes applicable to this problem. Namely, if we manage to choose a function $V(x, t)$ for which

$$V(0, t)=0,\ V(x, t)>0 \quad \text{for} \quad x\neq 0,\quad \dot{V}\leqslant 0$$

then for any solution we obtain

$$V(t, x)\leqslant V(t_0, x_0) \qquad (t_0\leqslant t\leqslant T) \tag{54}$$

whence an estimate on $\mathbf{x}(t)$ for a given estimate on $\mathbf{x}_0$ can be derived. For instance, if

$$V(x, t)=(\mathbf{A}(t)\mathbf{x}, \mathbf{x})e^{kt}$$

where $\mathbf{A}(t)$ is a symmetric positive definite matrix for every value of t, inequality (54) implies that

$$|\mathbf{x}(t)|\leqslant|\mathbf{x}_0|\sqrt{\frac{\lambda_{\max}(t_0)}{\lambda_{\min}(t)}}\exp\left[\frac{k(t_0-t)}{2}\right]$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are, respectively, the greatest and the least eigenvalues of the matrix $\mathbf{A}$ (prove this result!).

If the right members of the system are also subject to perturbation the inequality $\dot{V}\leqslant 0$ may turn out to hold only outside a neighbour-

hood, of the origin of the x-space, whose diameter depends on the given estimate for the perturbations of the right-hand members. The inequality (54) is then also guaranteed only until the solution enters this neighbourhood. For a linear system of type (3) it is also possible to make use of formula (1.12) assuming that the perturbation function $\mathbf{f}(t)$ satisfies the given inequality.

For the questions mentioned in this section the reader is referred to [59].

§ 4. NONLINEAR VIBRATIONS

1. Introduction. We shall begin with some well-known facts of the theory of oscillations of autonomous linear systems with one degree of freedom (e. g. see [107], Secs. XV.15, 17 and 18). The equation of free vibrations of such a system has the form

$$\ddot{y}+2h\dot{y}+\omega_0^2 y=0 \tag{1}$$

where $h\geqslant 0$ and $\omega_0>0$ are the parameters of the system. (In the case of a mass point suspended by a spring we have $2h=fm^{-1}$ and $\omega_0^2=km^{-1}$ where f is the coefficient of friction, m is the mass, k is the stiffness factor and y is the vertical coordinate.) The general solution of equation (1) is readily found:

$$y=e^{-ht}(C_1\cos\omega_1 t+C_2\sin\omega_1 t)\quad(\omega_1=\sqrt{\omega_0^2-h^2}) \tag{2}$$

if $h<\omega_0$ (for $h>0$ this formula describes **damped oscillations**), and

$$y=C_1\exp\left[-(h-\sqrt{h^2-\omega_0^2})\,t\right]+C_2\exp\left[-(h+\sqrt{h^2-\omega_0^2})\,t\right]$$

if $h>\omega_0$ (the so-called **aperiodic motion**). Based on formula (2), we call h a **damping factor**, the quantity h^{-1} being the **time constant** equal to a time interval during which the amplitude of oscillations decreases e times. (More precisely, one should speak here about the damping of the exponential factor in the sinusoid since the notion of the amplitude is only applicable, in the strict sense, to periodic functions to whose class function (2) does not belong; but nevertheless the term "amplitude" is usually applied in this broadened sense.) The dimensionless quantity $2\pi h\omega_1^{-1}$ termed the **logarithmic decrement** is equal to the logarithm of the ratio of two consecutive amplitudes corresponding to a time interval $\frac{2\pi}{\omega_1}$ (conditionally referred to as the **period of oscillations**) which is the period of the harmonic factors in (2). Finally, the ratio of two consecutive amplitudes is known as the (**damping**) **decrement**.

If there is no damping, that is $h=0$, solution (2) takes the simpler form

$$y=C_1\cos\omega_0 t+C_2\sin\omega_0 t=M\sin(\omega_0 t+\varphi_0)$$

where C_1 and C_2 (or M and φ_0) are arbitrary constants.

An important feature is that the frequency of vibrations ω_0 is independent of the amplitude M; as will be shown in Sec. 2, this property (known as isochronism) does not take place in the case of general nonlinear autonomous systems.

In the case of a harmonic external action the equation of forced oscillations is written in the form

$$\ddot{x} + 2h\dot{x} + \omega_0^2 x = A\cos\omega t \tag{3}$$

where A and ω are, respectively, the amplitude and the frequency of the external harmonic force. The general solution of equation (3) is (check it up!)

$$x = \frac{A}{(\omega^2 - \omega_0^2)^2 + 4h^2\omega^2}\left[-(\omega^2 - \omega_0^2)\cos\omega t + 2h\omega\sin\omega t\right] + \\ + e^{-ht}(C_1\cos\omega_1 t + C_2\sin\omega_1 t) \tag{4}$$

This solution is the result of the superposition of a harmonic vibration, whose frequency coincides with the frequency of the external action, and a free vibration. The former is uniquely determined by the right member of (3) while the latter is of a general form and involves two arbitrary constants dependent on the initial conditions. If $h > 0$ (the presence of dissipation) the second term in (4) is damped as t increases; therefore, practically, after a sufficiently long time only the first (harmonic) oscillation remains, its amplitude being

$$M = A\left[(\omega^2 - \omega_0^2)^2 + 4h^2\omega^2\right]^{-\frac{1}{2}}$$

The dimensionless coefficient $\frac{M\omega_0^2}{A}$ termed the **magnification (amplification) factor** or **gain (gain coefficient)** is expressed by the formula

$$\frac{M\omega_0^2}{A} = \left[\left(\left(\frac{\omega}{\omega_0}\right)^2 - 1\right)^2 + \right. \\ \left. + 4\left(\frac{h}{\omega_0}\right)^2\left(\frac{\omega}{\omega_0}\right)^2\right]^{-\frac{1}{2}} \tag{5}$$

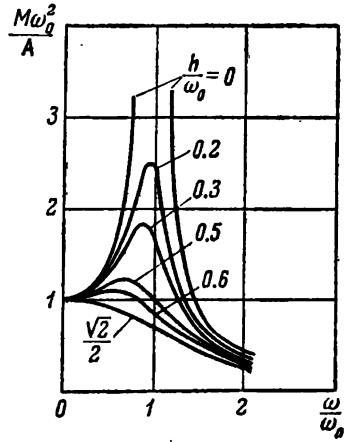


Fig. 178

It is a function of two dimensionless parameters $\frac{\omega}{\omega_0}$ and $\frac{h}{\omega_0}$, the former being the dimensionless excitation frequency and the latter the dimensionless dissipation coefficient. The dependence of the amplification factor on the first parameter for fixed values of the other parameter is shown in Fig. 178. Such graphs are termed **amplitude-frequency characteristics**.

For a fixed dissipation coefficient $r=\frac{h}{\omega_0}\leqslant\frac{\sqrt{2}}{2}$ the coefficient (5) attains its greatest value $(2r\sqrt{1-r^2})^{-1}$ for $\frac{\omega}{\omega_0}=\sqrt{1-2r^2}$ (check it up; also consider the case $r>\frac{\sqrt{2}}{2}$). For $\frac{h}{\omega_0}\to 0$ and $\frac{\omega}{\omega_0}\to 1$ the gain increases unlimitedly (what is the physical meaning of this fact?).

If there is no dissipation formula (4) is simplified:

$$x=\frac{A}{\omega_0^2-\omega^2}\cos\omega t+C_1\cos\omega_0 t+C_2\sin\omega_0 t \tag{6}$$

Formula (6) describes a quasiperiodic solution (Sec. 1.7) if the frequencies ω and ω_0 are incommensurable and a periodic one for the commensurable frequencies. In particular, for initial conditions

$$x(0)=0,\quad \dot{x}(0)=0$$

formula (6) yields the expression

$$x=\frac{A}{\omega_0^2-\omega^2}(\cos\omega t-\cos\omega_0 t) \tag{7}$$

Therefore if the excitation frequency ω coincides with the natural frequency ω_0 we have

$$x|_{\omega=\omega_0}=\lim_{\omega\to\omega_0}\frac{A}{\omega_0^2-\omega^2}(\cos\omega t-\cos\omega_0 t)=\frac{A}{2\omega_0}t\sin\omega_0 t \tag{8}$$

that is there appears a vibration whose amplitude grows indefinitely according to a linear law for $t\to\infty$; this means that resonance sets in. We can also say that in this case the magnification factor turns into infinity (see Fig. 178).

Solution (7) possesses some interesting features when ω is close to ω_0, that is $\omega=\omega_0+\delta$ where δ is relatively small. Then

$$x=\frac{2A}{\delta(\omega_0+\omega)}\sin\frac{\delta}{2}t\sin\frac{\omega_0+\omega}{2}t\approx\frac{A}{\delta\omega_0}\sin\frac{\delta}{2}t\sin\omega_0 t \tag{9}$$

For $0<t\ll\delta^{-1}$ the approximate relation $\frac{1}{\delta}\sin\frac{\delta}{2}t\approx\frac{t}{2}$ holds, and hence, as t increases, the amplitude grows like in the resonance phenomenon. But when t gains further increase the factor $\frac{1}{\delta}\sin\frac{\delta}{2}t$ oscillates according to a harmonic law with large amplitude δ^{-1} and small frequency $\frac{\delta}{2}$. The graph of the solution is therefore of the form shown in Fig. 179; this is known as **beating phenomenon** (the "splashes" occurring here with period $\frac{2\pi}{\delta}$ are called **beats**).

Let us also consider another case when a small external excitation is proportional to the sought-for solution; this situation is described by the equation

$$\ddot{x}+\omega_0^2 x=\alpha x \tag{10}$$

where $|\alpha|$ is small. For definiteness, let us take the initial conditions

$$x(0)=A,\ \dot{x}(0)=0 \tag{11}$$

The exact solution is readily obtained:

$$x=A\cos\sqrt{\omega_0^2-\alpha}\,t \tag{12}$$

(check up this expression!). Thus, such an excitation leads to a change of the frequency of vibrations, this fact being obvious

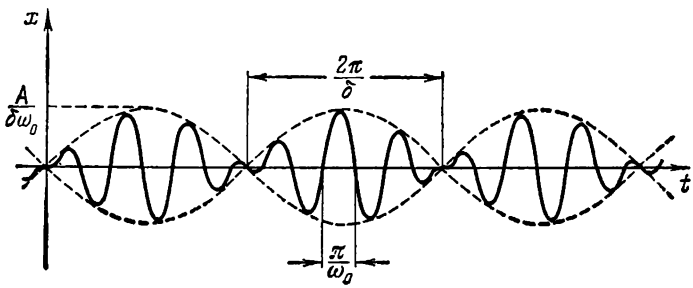


Fig. 179

since the introduction of an excitation of this type is equivalent to the corresponding variation of the stiffness factor.

Next suppose that we are going to apply to the problem (10), (11) the standard scheme of the small-parameter method. Then we write

$$x(t)=x_0(t)+x_1(t)\,\alpha+x_2(t)\,\alpha^2+\ldots \tag{13}$$

and substitute (13) into (10) and (11), which yields

$$\ddot{x}_0+\omega_0^2 x_0=0,\ x_0(0)=A,\ \dot{x}_0(0)=0 \quad \text{whence} \quad x_0(t)=A\cos\omega_0 t$$

the latter expression being the unperturbed solution. Furthermore, we obtain

$$\ddot{x}_1+\omega_0^2 x_1=x_0(t)=A\cos\omega_0 t,\ x_1(0)=0,\ \dot{x}_1(0)=0$$

whence follows that $x_1(t)$ is expressed by formula (8). Limiting ourselves to this approximation we receive the approximate solution

$$x(t)\approx A\cos\omega_0 t+\frac{A\alpha}{2\omega_0}\,t\sin\omega_0 t \tag{14}$$

including the correction of the first order of smallness due to the external excitation. The same result is obtained after the first

iteration if we apply to the problem (10), (11) the scheme of the iteration method.

For finite values of t we thus obtain a correct description of the phenomenon, but if we are interested in the behaviour of $x(t)$ on an infinite time interval, for instance, as $t \to \infty$, it is apparent that the "approximate" solution (14) is completely incorrect, its accuracy in describing the oscillations being even worse than that of the unperturbed solution $x_0(t)$. Indeed, even for small $|\alpha| \neq 0$ the second term on the right-hand side of (14) increases unboundedly, for $t \to \infty$, in its amplitude and therefore dominates the first term. (Such summands with amplitude growing according to a power law are called **secular terms**.) Formula (14) looks as if there were resonance in the considered system while the exact solution (12) shows that this is not the case. The addition of subsequent terms of expansion (13) even worsens the approximation; for instance, the expression $x_2(t) = -\frac{A}{8\omega^2} t^2 \cos \omega t + \frac{A}{8\omega^3} t \sin \omega t$ (let the reader verify that it corresponds to the second (formal) approximation of the small-parameter method) grows for $t \to \infty$ still faster than $x_1(t)$ (although on a finite time interval the accuracy is improved).

The same representation (13) can be derived directly by expanding exact solution (12) in powers of α. Actually, putting $\sqrt{\omega_0^2 - \alpha} = \omega_0 - \beta$ we receive

$$\beta = \omega_0 - \omega_0 \left(1 - \frac{\alpha}{\omega_0^2}\right)^{\frac{1}{2}} = \frac{\alpha}{2\omega_0} + \frac{\alpha^2}{8\omega_0^3} + \frac{\alpha^3}{16\omega_0^5} + \ldots \qquad (15)$$

Furthermore, we have

$$x(t) = A \cos(\omega_0 - \beta) t = A \cos \omega_0 t \cos \beta t + A \sin \omega_0 t \sin \beta t =$$
$$= A \cos \omega_0 t \cdot \left(1 - \frac{\beta^2 t^2}{2!} + \frac{\beta^4 t^4}{4!} - \ldots\right) + A \sin \omega_0 t \cdot \left(\beta t - \frac{\beta^3 t^3}{3!} + \ldots\right)$$

If now we substitute expansion (15) into the right-hand side of the last formula and combine the terms with like powers of α we obtain (13). It therefore becomes clear why the secular terms creating an illusion of resonance appear: partial sums of Maclaurin's series for cos s and sin s give a good approximation to these functions on every finite interval of s but are completely inapplicable when $s \to \infty$ (analyse this situation!).

In more complicated problems it is impossible to obtain an explicit finite expression for the exact solution, and then it becomes more difficult to recognize the illusion of resonance created by secular terms. Such terms usually occur in the investigation of an autonomous system by means of a perturbation scheme when a vibration with an incorrectly determined frequency is taken as the zeroth approximation (because, given an arbitrarily small ε, $|\varepsilon| > 0$, the difference $\cos(\omega_0 + \varepsilon) t - \cos \omega_0 t$ is not small for $t \to \infty$).

The properties of autonomous linear vibration systems with a finite number $n > 1$ of degrees of freedom are established on the basis of the considerations given in Secs. 3.6 and 1.6 and 7. Here we shall dwell only on the notion of the amplitude-frequency characteristic of such a system. If the equation of forced oscillations of a system of that kind under a harmonic external action with frequency ω is written in the form

$$\mathbf{a}\ddot{\xi} + \mathbf{c}\dot{\xi} + \mathbf{b}\xi = \mathbf{d}e^{i\omega t} \tag{16}$$

where $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are square matrices, $\xi(t)$ is the sought-for vector function and $\mathbf{d}$ is a given constant vector, its solution having the form of a harmonic with the same frequency is equal to

$$\xi = (-\omega^2\mathbf{a} + i\omega\mathbf{c} + \mathbf{b})^{-1}\,\mathbf{d}e^{i\omega t}$$

To obtain the general solution we must add to this function the general expression for free vibrations of the system which depend on the initial conditions. If there is total dissipation this additional term (all its coordinates) is damped as $t \to \infty$. Consequently, in this case the role of the amplitude-frequency characteristic is played by the matrix

$$\mathbf{Z}(\omega) = (-\omega^2\mathbf{a} + i\omega\mathbf{c} + \mathbf{b})^{-1} \tag{17}$$

If (16) is a Lagrange's equation for the system under investigation every element z_{jk} of the matrix $\mathbf{Z} = (z_{jk})$ is a complex magnification factor characterizing the response of the jth coordinate to a harmonic action with frequency ω applied to the kth coordinate. Thus, the matrix (17) characterizes the interrelations between the elements of the system with respect to a harmonic excitation.

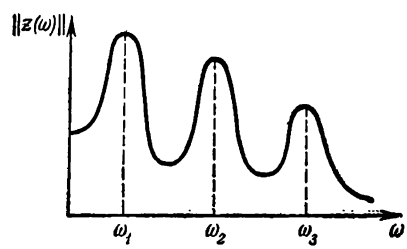


Fig. 180

Instead of the complete characteristic (17) the simplified amplitude-frequency characteristic $\|\mathbf{Z}(\omega)\|$ is sometimes used (the norm of the matrix is understood as the sum of the moduli of its elements). This is a real scalar function of the real argument ω, and therefore we can plot its graph (see Fig. 180). It is evident that if the norm of the matrix $\mathbf{Z}(\omega)$ is small all the gain factors are small for the given ω, which means that a harmonic excitation with this frequency does not notably affect the system. But if the norm is large (in Fig. 180 this is the case for $\omega = \omega_1$, $\omega = \omega_2$ and $\omega = \omega_3$) the effect is strong (more precisely, a harmonic action with this frequency can be chosen in such a way that at least one element, component, of the system is in a vibration with a large amplitude). For a system without dissipation

an excitation with the natural frequency ω_0 leads to resonance, that is $\| \hat{\mathbf{Z}}(\omega_0) \| = \infty$.

As an example, consider the system of two equations

$$\ddot{x}_1 + \omega_{01}^2 x_1 = r(x_2 - x_1),$$
$$\ddot{x}_2 + \omega_{02}^2 x_2 = r(x_1 - x_2) \quad (r > 0)$$

describing the vibrations of a system of two oscillators connected by means of an elastic coupling (see Sec. 3.6). In this case we obtain (check up the calculations!)

$$\hat{\mathbf{Z}}(\omega) = [\omega^4 - (\omega_{01}^2 + \omega_{02}^2 + 2r)\omega^2 + \omega_{01}^2\omega_{02}^2 + r(\omega_{01}^2 + \omega_{02}^2)]^{-1} \times$$
$$\times \begin{pmatrix} -\omega^2 + \omega_{02}^2 + r & r \\ r & -\omega^2 + \omega_{01}^2 + r \end{pmatrix}$$

In particular,

$$z_{21}(\omega_{01}) = (\omega_{02}^2 - \omega_{01}^2)^{-1}$$

which means that if the first oscillator is excited by a harmonic whose frequency coincides with its natural frequency then, the closer the natural frequencies of the oscillators, the stronger the response of the second oscillator.

In the next section we shall systematically study nonlinear vibrations which possess some peculiarities even in the case of one degree of freedom. It is expedient to point out some of them here. Firstly, the superposition principle is inapplicable to nonlinear systems (for example, this means that a solution cannot be multiplied by a constant, etc.). This results in nonisochronism of a free vibration of an autonomous conservative nonlinear system, that is in the dependence of the frequency of such a vibration on its amplitude. Secondly, such a system does not possess a (constant) natural frequency, and therefore, when the system is subjected to a harmonic excitation with a constant frequency the (ordinary) resonance does not set in. Thirdly, as will be shown, in a system of that kind there may appear a new phenomenon which is impossible, in principle, in a linear system; in some circumstances there arises a forced vibration whose period is an integral multiple of the period of the external excitation (a so-called **subharmonic vibration**). Another essentially new feature is that in a nonlinear autonomous nonconservative system autovibrations are possible (see Sec. 2.5 where this phenomenon was mentioned). Finally, the inapplicability of the superposition principle may result in the appearance, in a nonautonomous nonlinear system, of **parametric vibrations** which are generated for certain values of the amplitude by periodic variations of the parameters of the system. If this amplitude becomes infinite there is parametric resonance in the system (Sec. 1.4). Finally, the investigation of nonlinear vibration

systems with several degrees of freedom is connected with complications due to the impossibility of passing to independent oscillations of normal coordinates.

There is extensive literature devoted to the theory of vibrations and, in particular, to nonlinear vibrations; for instance, see [4], [6], [12], [15], [22], [26], [51], [55], [61], [62], [91], [104], [108], [109], [124], [125], [127] and [128].

2. Free Oscillations of a Conservative Autonomous System with One Degree of Freedom. Here we shall deal with the equation

$$\ddot{x}+f(x)=0. \tag{18}$$

in which the sign of the function $f(x)$ coincides with that of x. Equation (18) possesses the first integral

$$V(x,\dot{x})\equiv\frac{1}{2}\dot{x}^2+\int_0^x f(\xi)\,d\xi=\text{const} \tag{19}$$

(check it up!) and is therefore integrable by quadratures.

The left member of (19) having a minimum for $x=\dot{x}=0$, the solutions are periodic for sufficiently small initial values of x and $\dot{x}$ (cf. Sec. 3.6). For a function $f(x)$ which is not odd the maximum amplitude $A_{\max}$, that is the maximum positive deviation of x from the equilibrium solution $x=0$ (satisfying (18) because $f(0)=0$), and the minimum amplitude $A_{\min}$ (understood as the absolute value of the minimum deviation) are different, in the general case. At the time moments when these maximum and minimum deviations are achieved we have $\dot{x}=0$, and therefore $A_{\max}$ and $A_{\min}$ are connected by the relation

$$-\int_{-A_{\min}}^{0} f(x)\,dx=\int_0^{A_{\max}} f(x)\,dx$$

If $f(x)$ is an odd function we obviously have $A_{\min}\equiv A_{\max}$.

Now let us investigate the connection between the amplitude and the period of oscillations. To this end we put $x_0=A_{\max}$ and $\dot{x}_0=0$ for $t=0$ and substitute these values into (19), which yields the relation

$$\frac{1}{2}\dot{x}^2=\int_0^x f(\xi)\,d\xi=\int_0^{A_{\max}} f(\xi)\,d\xi$$

whence

$$\dot{x}=\pm\left(2\int_x^{A_{\max}} f(\xi)\,d\xi\right)^{\frac{1}{2}}$$

Consequently, it takes the coordinate x time

$$\int_0^{A_{\max}} \left(2 \int_x^{A_{\max}} f(\xi)\, d\xi \right)^{-\frac{1}{2}} dx$$

to pass from the value $A_{\max}$ to zero.

Carrying out the same calculations for $x < 0$ and adding together the results we find that the period of vibrations is

$$T = 2 \int_{-A_{\min}}^{0} \left(-2 \int_{-A_{\min}}^{x} f(\xi)\, d\xi \right)^{-\frac{1}{2}} dx + 2 \int_0^{A_{\max}} \left(2 \int_x^{A_{\max}} f(\xi)\, d\xi \right)^{-\frac{1}{2}} dx \quad (20)$$

For instance, putting $f(x) = \alpha |x|^k \operatorname{sgn} x$ $(\alpha > 0,\ -1 < k < \infty)$ we obtain $A_{\max} = A_{\min} = A$ and

$$T = 4 \int_0^A \left(2 \int_x^A \alpha \xi^k\, d\xi \right)^{-\frac{1}{2}} dx = \sqrt{\frac{8(k+1)}{\alpha}} \int_0^A \left(A^{k+1} - x^{k+1} \right)^{-\frac{1}{2}} dx =$$

$$= \sqrt{\frac{8}{\alpha (k+1) A^{k-1}}} \int_0^1 (1-s)^{-\frac{1}{2}} s^{-\frac{k}{k+1}}\, ds =$$

$$= \sqrt{\frac{8\pi}{\alpha(k+1)}} \frac{\Gamma\left(\frac{1}{k+1}\right)}{\Gamma\left(\frac{1}{2} + \frac{1}{k+1}\right)} A^{\frac{1-k}{2}}$$

(we have made the substitution $x = As^{\frac{1}{k-1}}$) where Γ designates the gamma function (e.g. see [107], Secs. XIV.17 and 18). If $k = 1$ we deal with the linear case in which the amplitude and the period are independent; in all the other cases the period depends on the amplitude. For $k > 1$ the period decreases when the amplitude grows. This property is irrelevant to the concrete form of the function $f(x)$ and can be deduced, in the general case, from the fact that the graph of the function is convex downward for $x > 0$ and convex upward for $x < 0$; physically, this is accounted for by the dominant role of the elastic forces, as the amplitude grows, in comparison with the inertia forces (analyse this argument). The function $f(x)$ is the **restoring force characteristic,** and in the case of this type of convexity of its graph it is said to be **hard.** Similarly, for $k < 1$ we have a **soft** characteristic with opposite direction of convexity; in this case the period of oscillations grows together with the amplitude.

In approximate calculations it is sometimes possible to replace $f(x)$, on the basis of some physical considerations, by an appropriate linear function, after which the period is easily determined. It turns out that for an odd function $f(x)$ one of the most

accurate ways is to replace $f(x)$ by kx where k is determined by the minimization condition for the integral

$$\int_0^A [f(x)-kx]^2 x^2 dx$$

This leads to the approximate formula

$$T \approx 2\pi \left(\frac{5}{A^5} \int_0^A f(x) x^3 dx \right)^{-\frac{1}{2}}$$

for the period of vibrations, which is much simpler than (20).

To visualize the solutions of equation (18) it is convenient to pass to the phase plane (Secs. 2.3-5) of the equivalent system of the equations of the first order

$$\dot{x}=y, \quad \dot{y}=-f(x) \tag{21}$$

System (21) being canonical (write down its Hamiltonian function), the area is preserved under the shift along its phase trajectories (see Sec. 2.12). Consequently, the phase diagram cannot contain rest points attracting or repulsing trajectories (that is nodal and focal points) and limit cycles possessing the same property. In the construction of the phase diagram of system (21) it should be taken into account that all the trajectories go to the right in the upper half-plane $y>0$ (the x-axis is supposed to be horizontal) and to the left in the lower half-plane $y<0$. At all the points of intersection of the trajectories with the x-axis except the rest points (which can only be located on this axis) the trajectories are orthogonal to the axis (such properties of the phase diagram are characteristic of a more general equation $\ddot{x}+f(x, \dot{x})=0$ as well).

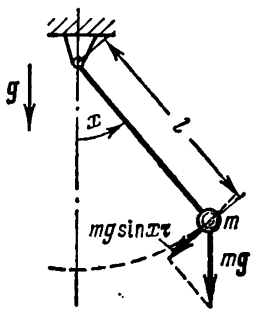


Fig. 181

We shall proceed to investigate the example of a simple pendulum (a material point suspended by an absolutely rigid inextensible weightless rod in a homogeneous field of gravitation with constant acceleration of gravity g; see Fig. 181). Using the notation in Fig. 181 we write

$$ml^2\ddot{x} = -mg \sin x \cdot l$$

whence

$$\ddot{x} + \frac{g}{l} \sin x = 0$$

(let the reader verify the coincidence of the dimensions of the terms of the equation). The corresponding system of the first order is of the form

$$\dot{x}=y, \quad \dot{y}=-\frac{g}{l}\sin x$$

its phase diagram being shown in Fig. 182 (check it!). The variable x being an angle, the diagram has a periodic structure (along the x-axis) with period 2π, and the phase space is topologically equivalent to a cylinder (see Sec. 2.11). The rest points lie on the x-axis and are stable centres whose abscissas are even integer

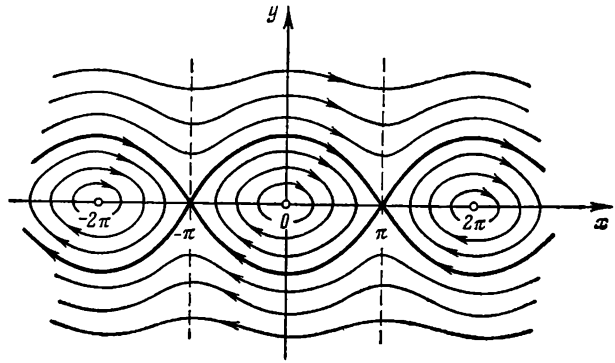


Fig. 182

multiples of π and unstable saddle points whose abscissas are odd integer multiples of π; the saddle points correspond to unstable equilibrium positions of the pendulum. Integral (19) implies that the equation of the trajectories can be written as

$$y^2=C+2\frac{g}{l}\cos x$$

where C is an arbitrary constant. In particular, the equation of the separatrix having two symmetric branches with respect to the x-axis and passing through the saddle points is obtained from the last equation if we put $x=\pi$ and $y=0$ to obtain the value $C=2\frac{g}{l}$, which results in the equation

$$y=\pm\sqrt{2\frac{g}{l}(1+\cos x)}=\pm 2\sqrt{\frac{g}{l}}\cos\frac{x}{2}$$

The separatrix divides the strip $-\pi\leqslant x\leqslant\pi$ (Fig. 182) into three parts. The part containing the x-axis is filled with cycles. In the regimes corresponding to these cycles the pendulum is in a vibrational motion (referred to as a **libration**) about the point $x=0$ of

stable equilibrium. The upper part is filled with the trajectories which are closed when considered on the phase cylinder and nonclosed in the phase plane. Every such trajectory corresponds to a **rotational motion** of the pendulum about the point of suspension with indefinitely increasing angle x. The lower part corresponds to the rotational motions with the opposite direction of rotation. (What kind of motion corresponds to the separatrix itself?) The connection between the period and the amplitude of a libration motion is computed by formula (20) and can be expressed by means of a complete elliptic integral of the first kind (see Sec. II.3.14), which we leave to the reader.

It should be noted that in this example every periodic motion is Lyapunov unstable. Indeed, the period of oscillations being dependent on the amplitude, every infinitesimal shift of the moving point to a neighbouring trajectory results in a finite deviation between the points (representing the motions in the phase plane) for infinitely increasing t (these points again become close after a certain time, then deviate repeatedly, etc.). On the other hand, the trajectories of motion of the representing points will remain close to each other all the time; in such a case we speak about **orbital stability.**

In theoretical physics libration and rotation are investigated by means of a special coordinate system distinct from x, $\dot{x}$. We shall demonstrate this technique by the example of an autonomous conservative system with one degree of freedom whose Hamiltonian function is

$$H(q,\ p)=\frac{p^2}{2m}+U(q)$$

where the parameter $m>0$ may be dependent on q. According to Sec. VI.3.7, the Hamilton-Jacobi equation takes, in this case, the form

$$\frac{1}{2m}\left(\frac{dW}{dq}\right)^2+U(q)=h$$

whence

$$W=\int \sqrt{2m\,[h-U(q)]}\,dq+\text{const} \tag{22}$$

The substitution of an expression $h=h(P)$ into the last integral results in a function $W(q,\ P)$ which can be taken as a generating function of the canonical transformation

$$p=\frac{\partial W}{\partial q},\quad Q=\frac{\partial W}{\partial P},\quad \tilde{H}=h(P) \tag{23}$$

After the transformation the law of motion takes the form (see Sec. VI.3.7)

$$P=\text{const},\quad Q=h'(P)\,t+\text{const}$$

Formulas (22) and (23) show that h is the total energy of the system.

For periodic motions in the vicinity of a stable equilibrium position the above-mentioned special system of variables is obtained if we take as P the so-called **action variable**

$$J=\frac{1}{2\pi}\oint p\,dq \tag{24}$$

which is equal, to within the factor $\frac{1}{2\pi}$, to the area bounded by the closed trajectory on the phase plane q, p. (Check that the increment of the Hamiltonian action integral along one cycle of period T_0 is equal to $2\pi J-hT_0$; this accounts for the term "the action variable".) The quantity J increasing together with h, we can speak about the inverse function $h(J)$ generating a canonical transformation of type (23). The quantity $\theta=\frac{\partial W}{\partial J}$ which is the canonical conjugate variable to J is termed the **angular variable.** Its increment Δ for one cycle is readily found:

$$\Delta\theta=\oint\frac{\partial\theta}{\partial q}\Big|_{J=\text{const}}dq=\oint\frac{\partial^2W}{\partial J\,\partial q}dq=\frac{\partial}{\partial J}\oint\frac{\partial W}{\partial q}dq=\frac{\partial}{\partial J}\oint p\,dq=2\pi$$

In the new variables the law of motion takes the form

$$\theta=h'(J)\,t+\theta_0=\frac{2\pi}{T_0}t+\theta_0,\quad J=\text{const}$$

As an exercise, let the reader show that for $m=\text{const}$ and a linear restoring force, that is for $U=\frac{k}{2}q^2$, these formulas yield

$$J=\sqrt{\frac{m}{k}}\,h,\quad \theta=\arcsin\frac{\sqrt[4]{mk}\,q}{\sqrt{2J}}$$

$$q=\frac{\sqrt{2J}}{\sqrt[4]{mk}}\sin\theta,\quad p=\sqrt[4]{mk}\,\sqrt{2J}\cos\theta$$

For a rotational motion integral (24) is taken over a period of the function $p(q)$, that is over the cycle (on the phase cylinder) corresponding to the rotation, while all the other computations remain the same as in the case of libration.

In the presence of dissipation the phase diagram cannot contain cycles, the rest points corresponding to the stable equilibrium states turn into focal points when the dissipation is small and into nodal points when it is large (why?). Of course in this case the system is no longer conservative. The phase diagram for a simple pendulum with small dissipation is depicted in Fig 183. (Let the reader make the construction for large dissipation.) For any initial data, even for an arbitrarily large initial velocity, the pendulum,

possibly after a finite number of rotations, passes to a damped libration. (What is the role of the separatrix in this situation?)

In connection with these questions we mention another useful notion which has received extensive study in recent years. Consider a system of the general form

$$\dot{\mathbf{x}} = \mathbf{f}(x, t) \tag{25}$$

A k-dimensional ($1 \leqslant k \leqslant n$) manifold ($S$) of the $(n+1)$-dimensional space of the variables x_j, t ($j = 1, 2, \ldots, n$) is called an **integral manifold** of system (25) if it is entirely composed of the graphs

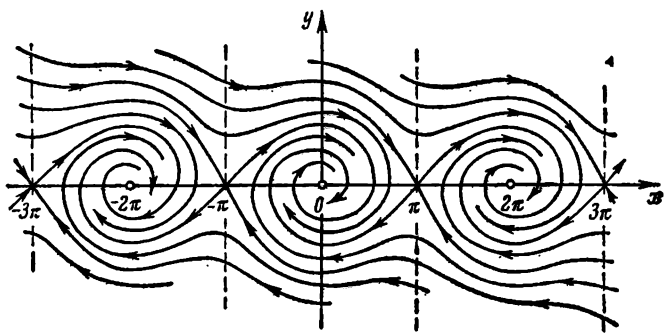


Fig. 183

of solutions (integral curves), that is if $(x_0, t_0) \in (S)$ implies $(x(t, t_0, x_0), t) \in (S)$ for all t where $x(t, t_0, x_0)$ is the solution of system (25) satisfying the initial condition $x|_{t=t_0} = x_0$ ($x = (x_1, \ldots, x_n)$ and $x_0 = (x_{01}, \ldots, x_{0n})$). Every integral curve is of course a one-dimensional integral manifold. If system (25) is autonomous its solutions are invariant with respect to time shift $t \to t + C$, and therefore every trajectory in the phase space x gives rise to a two-dimensional (cylindrical) integral manifold in the space x, t, its directrix being the trajectory and its generators being parallel to the t-axis. The investigation of integral manifolds sometimes enables us to abstract from individual solutions and to study basic global properties of the solutions. (In this connection remember the notion of a first integral (e.g. see [107], Sec. XV.13). It is clear that every first integral specifies a stratification of the x, t-space into n-dimensional integral manifolds.) Extensive research has also been devoted to the dependence of an integral manifold on parameters which may be involved in system (25) and to the Lyapunov stability of the section of an integral manifold by a hyperplane $t = \text{const}$.

3. Forced Oscillations of a Weakly Nonlinear System. Basic Case. Let us consider the equation

$$\ddot{x}+2h\dot{x}+\omega_0^2 x = A\cos\omega t + f_1(x,\ \dot{x},\ t)\,\alpha + f_2(x,\ \dot{x},\ t)\,\alpha^2 + \ldots \quad (26)$$

where α is a small parameter and all the functions f_j are periodic in t with period $T=\frac{2\pi}{\omega}$ and can be expanded into Taylor's series in powers of x and $\dot{x}$. This is an equation of type (3) perturbed by a small nonlinear term. We shall suppose that $A\neq 0$; for the case $h=0$ we shall also assume that $\omega\neq\omega_0$ (this excludes singular cases which will be treated separately in Sec. 4).

The coefficients h, ω_0^2 and A in equation (3) can also be given small perturbations but this does not lead to a significant generalization since if, for instance, we take, instead of h, an expansion of the form $h+h_1\alpha+h_2\alpha^2+\ldots$, the terms $h_j\alpha^j$ $(j=1,\ 2,\ \ldots)$ can be combined with $f_j\alpha^j$. But if there is a perturbation of the excitation frequency, that is ω is everywhere replaced by $\omega+\omega_1\alpha+$ $+\omega_2\alpha^2+\ldots$ $\left(\text{the period } T=\frac{2\pi}{\omega} \text{ then also becomes dependent on } \alpha\right)$, the cosine and the functions f_j must not be expanded in powers of α because this would lead to the appearance of secular terms (see Sec. 1). In this case it is advisable to consider the whole expression $\omega+\omega_1\alpha+\omega_2\alpha^2+\ldots$ as a new frequency.

According to the small-parameter method, we shall seek the solution of equation (26) in the form

$$x = x_0(t)+x_1(t)\,\alpha+x_2(t)\,\alpha^2+\ldots \quad (27)$$

Substituting (27) into (26) and equating the coefficients in the same powers of α we arrive at the relations

$$\ddot{x}_0+2h\dot{x}_0+\omega_0^2 x_0 = A\cos\omega t \quad (28)$$

$$\ddot{x}_1+2h\dot{x}_1+\omega_0^2 x_1 = f_1(x_0,\ \dot{x}_0,\ t) \quad (29)$$

$$\ddot{x}_2+2h\dot{x}_2+\omega_0^2 x_2 = \left[\frac{\partial}{\partial\alpha} f_1(x_0+\alpha x_1,\ \dot{x}_0+\alpha\dot{x}_1,\ t)\right]_{\alpha=0} + f_2(x_0,\ \dot{x}_0,\ t) \quad (30)$$

and so on. Equation (28) coincides with equation (3), and hence $x_0(t)$ is expressed by formula (4). We shall only look for a periodic solution with period T which is most interesting. Then

$$x_0(t) = \frac{A}{(\omega^2-\omega_0^2)^2+4h^2\omega^2}\left[-(\omega^2-\omega_0^2)\cos\omega t+2h\omega\sin\omega t\right] \quad (3)$$

Substituting this expression into the right-hand side of (29) we obtain a periodic function of t with period T which can be expanded into Fourier's series with respect to the functions $\cos n\omega t$ and

$\sin n\omega t$ $(n=1, 2, \ldots)$. By analogy with (31), we then find $x_1(t)$ in the form of a series in the same functions, that is as a periodic function of t with period T. After that we repeat the procedure for equation (30) and so on.

Let us apply this scheme to Duffing's equation

$$\ddot{x}+\omega_0^2 x = A\cos\omega t+\alpha x^3 \quad (\omega \neq \omega_0) \tag{32}$$

Here we have

$$x_0(t)=\frac{A}{\omega_0^2-\omega^2}\cos\omega t$$

and therefore equation (29) turns into

$$\ddot{x}_1+\omega_0^2 x_1=\left(\frac{A}{\omega_0^2-\omega^2}\cos\omega t\right)^3=\frac{A^3}{4(\omega_0^2-\omega^2)^3}(3\cos\omega t+\cos 3\omega t)$$

whence

$$x_1(t)=\frac{A^3}{4(\omega_0^2-\omega^2)^3}\left[\frac{3}{\omega_0^2-\omega^2}\cos\omega t+\frac{1}{\omega_0^2-(3\omega)^2}\cos 3\omega t\right] \tag{33}$$

(check up this expression!). The function $x_2(t)$ will be a linear combination of the functions $\cos\omega t$, $\cos 3\omega t$ and $\cos 5\omega t$, etc. Thus, the periodic solution of equation (32) is obtained as a superposition of a fundamental harmonic with frequency ω and higher (upper) harmonics with frequencies 3ω, 5ω, etc., whose amplitudes are of the orders of α, α^2, etc.

The same technique is applicable when the right-hand side of (26) contains, instead of $A\cos\omega t$, an arbitrary periodic function $f(t)$ with period $T=\frac{2\pi}{\omega}$. After such a right member of (28) is expanded into Fourier's series the construction of the solution is completely similar to the above. If $f(t)\equiv 0$ (that is $A=0$) we can make the change $x=\alpha y$, cancel equation (26) by α and expand all the terms $f_j(\alpha y, \alpha\dot{y}, t)$ in powers of α; if $f_1(0, 0, t)\equiv 0$ this transformation can be repeated. It should be noted that formulas (33) and similar formulas for equations with nonlinear terms of a more general type indicate that here not only the case $\omega=\omega_0$ is singular but also the cases $\omega=\frac{\omega_0}{2}$, $\frac{\omega_0}{3}$, etc.

For systems with a greater number of degrees of freedom, in the basic case, that is when for $\alpha=0$ none of the natural frequencies is an integer multiple of the excitation frequency, this scheme is applied in a completely similar way.

In practical applications we often deal with concrete numerical examples for which it is known, intuitively, that the nonlinearity involved is not too strong. In these circumstances one of the relatively small numerical coefficients generating nonlinearity can be taken as a "small parameter" (although it is constant!). Denoting

it by α we then can apply the above technique (e.g. cf. [107], Sec. XV.27). It is sometimes advisable to introduce an artificial factor α as a coefficient in the group of all nonlinear terms, apply the small-parameter scheme and then put $\alpha = 1$ in the ultimate result. Usually we limit ourselves to a finite number of terms of the expansion obtained; the actual accuracy of such an approximation can be tested by means of numerical solution of some test examples, with typical or extremal values of the parameters, on an electronic digital computer or on an analogue computer. These remarks also refer to the other modifications of the small-parameter method.

4. Singular Cases. Now let us consider the case when the excitation frequency in a weakly nonlinear system with one degree of freedom is close to the natural frequency, that is when $h = 0$ and $\omega = \omega_0$ in equation (26). We shall look for a periodic solution of period T which remains finite for $\alpha = 0$. Then we must have $A = 0$ in equation (26), and thus the equation under consideration is taken in the form

$$\ddot{x} + \omega_0^2 x = f_1(x, \dot{x}, t)\alpha + f_2(x, \dot{x}, t)\alpha^2 + \ldots \tag{34}$$

where all the functions f_j $(j = 1, 2, \ldots)$ are T_0-periodic in t $\left(T_0 = \frac{2\pi}{\omega_0}\right)$.

We shall use the same form (27) of the expansion. Then $\ddot{x}_0 + \omega_0^2 x_0 = 0$, that is

$$x_0(t) = a\cos\omega_0 t + b\sin\omega_0 t \tag{35}$$

where the constants a and b are yet unknown. This leads (see (29)) to the equation

$$\ddot{x}_1 + \omega_0^2 x_1 = f_1(a\cos\omega_0 t + b\sin\omega_0 t, -a\omega_0\sin\omega_0 t + b\omega_0\cos\omega_0 t, t) \tag{36}$$

Our task is to construct a periodic function $x_1(t)$, and therefore the expansion of the right-hand side of (36) into Fourier's series must not involve $\cos\omega_0 t$ and $\sin\omega_0 t$ (why?), that is the conditions

$$\left.\begin{aligned} &\int_0^{\frac{2\pi}{\omega_0}} f_1(a\cos\omega_0 t + b\sin\omega_0 t, -a\omega_0\sin\omega_0 t + b\omega_0\cos\omega_0 t, t)\times \\ &\qquad \times\cos\omega_0 t\, dt = 0 \\ &\int_0^{\frac{2\pi}{\omega_0}} f_1(a\cos\omega_0 t + b\sin\omega_0 t, -a\omega_0\sin\omega_0 t + b\omega_0\cos\omega_0 t, t)\times \\ &\qquad \times\sin\omega_0 t\, dt = 0 \end{aligned}\right\} \tag{37}$$

must be fulfilled. This is a system of the so-called **determining equations** which specify the possible values of a and b. This sys-

tem being, in the general case, nonlinear, the typical situation is when there are a finite number of such values (or none).

After certain values of a and b have been found from (37) we obtain the general solution of equation (36) in the form

$$x_1(t) = \varphi_1(t) + a_1 \cos \omega_0 t + b_1 \sin \omega_0 t \tag{38}$$

where $\varphi_1(t)$ is a concrete, completely determined, periodic function of t with period T_0 and a_1, b_1 are new arbitrary constants. The expressions of $x_0(t)$ and $x_1(t)$ are now substituted into equation (30) with the unknown function $x_2(t)$. The condition that $x_2(t)$ must not include resonant terms yields two determining equations of the first degree for a_1 and b_1 from which these constants are found. Next we form the general solution of equation (30) involving two new arbitrary parameters and so on.

It should be noted that in the case considered above a small external action generates a periodic vibration with a finite, not small, amplitude. This phenomenon can naturally be expected since the system is close to a resonance state. At the same time, the presence of a small dissipation (due to the derivative $\dot{x}$ entering into the right-hand side) and of a small nonlinearity resulting in a change of the natural frequency when the amplitude varies prevents this forced vibration from growing unlimitedly. (To elucidate what has been said, let the reader carry out all the calculations for the simple equation $\ddot{x} + \omega_0^2 x = -\alpha h \dot{x} + \alpha A \cos \omega_0 t$ belonging to this type.)

If the right-hand side of (34) contains an additional periodic function $f_0(t)$ of period T_0 whose Fourier's expansion does not involve $\cos \omega_0 t$ and $\sin \omega_0 t$, the problem is reduced to the above by means of the substitution $x = y + \psi(t)$ where the function $\psi(t)$ is an arbitrary (concrete) solution of the equation $\ddot{x} + \omega_0^2 x = f_0(t)$.

As an example, let us take a modification of Duffing's equation involving a small damping:

$$\ddot{x} + \alpha h \dot{x} + \omega_0^2 x = \alpha A \cos \omega t + \alpha k x^3 \qquad (h > 0,\ \alpha > 0)$$

Let the excitation frequency ω be close to ω_0 and have the expansion

$$\omega = \omega_0 + \omega_1 \alpha + \omega_2 \alpha^2 + \ldots \tag{39}$$

According to what has been said at the beginning of Sec. 3, we take expression (39) as a new frequency and make the change

$$\omega_0^2 = [\omega - (\omega_1 \alpha + \omega_2 \alpha^2 + \ldots)]^2 = \omega^2 - 2\omega\omega_1 \alpha + (\omega_1^2 - 2\omega\omega_2)\alpha^2 + \ldots$$

before expanding the solution in powers of α. This results in

$$\ddot{x} + \omega^2 x = (-h\dot{x} + A \cos \omega t + kx^3 + 2\omega\omega_1 x)\alpha - (\omega_1^2 - 2\omega\omega_2) x\alpha^2 - \ldots$$

Now, substituting the expansion $x = x_0(t) + x_1(t)\alpha + \ldots$ into the last equation we obtain, after the coefficients in like power of α have been equated, the relations

$$\ddot{x}_0 + \omega^2 x_0 = 0$$

$$\ddot{x}_1 + \omega^2 x_1 = -h\dot{x}_0 + A\cos\omega t + kx_0^3 + 2\omega\omega_1 x_0 \tag{40}$$

etc. The general solution of the first equation is of form (35) with ω_0 replaced by ω. Substituting this general solution into the right-hand side of (40) and writing down the conditions providing the absence of resonance we arrive, after some computations, at the determining equations

$$\left.\begin{aligned} 2\omega\omega_1 a - h\omega b + A + \tfrac{3}{4}k(a^3 + ab^2) = 0 \\ 2\omega\omega_1 b + h\omega a + \tfrac{3}{4}k(a^2 b + b^3) = 0 \end{aligned}\right\} \tag{41}$$

To investigate this system it is convenient to introduce the quantity $m = \sqrt{a^2 + b^2}$ which is equal to the limiting amplitude of oscillations for $\alpha \to 0$. From (41) we then obtain

$$\left(2\omega\omega_1 + \tfrac{3}{4}km^2\right)a - h\omega b = -A, \quad \left(2\omega\omega_1 + \tfrac{3}{4}km^2\right)b + h\omega a = 0$$

Squaring these relations and adding them together we receive the equation for the amplitude m:

$$\left(2\omega\omega_1 + \tfrac{3}{4}km^2\right)^2 m^2 + h^2\omega^2 m^2 = A^2 \tag{42}$$

To simplify the investigation of this equation we introduce the quantities

$$\mu = \frac{m}{A}\omega^2 s \quad \text{and} \quad \lambda = \frac{\omega_1}{\omega} r$$

which are, respectively, the dimensionless amplitude and the dimensionless increment of the frequency. The dimensionless coefficients s and r whose values are specified below are introduced in order to reduce the number of parameters (draw attention to this technique which is of general significance and is used in many problems). Instead of (42) we then obtain the equation

$$\tfrac{9}{16}k^2A^4\omega^{-12}s^{-6}\left(\tfrac{8}{3}s^2\omega^6(rkA^2)^{-1}\lambda + \mu^2\right)^2\mu^2 + h^2\omega^{-2}s^{-2}\mu^2 = 1$$

(check all the calculations!). Hence, if we put

$$s = \left(\tfrac{3}{4}kA^2\right)^{\frac{1}{3}}\omega^{-2}, \quad r = \tfrac{8}{3}s^2\omega^6(kA^2)^{-1} = \left(\tfrac{3}{32}kA^2\right)^{-\frac{1}{3}}\omega^2$$

and $q = h^2\omega^{-2}s^{-2} = h^2\left(\frac{3}{4}kA^2\right)^{-\frac{2}{3}}\omega^2$ equation (42) reduces to

$$(\lambda+\mu^2)^2\mu^2 + q\mu^2 = 1, \text{ that is } (\lambda+\mu^2)^2 + q = \frac{1}{\mu^2}$$

The graphical solution of this final equation with some given λ and q is shown in Fig. 184. Since the equation is of the third

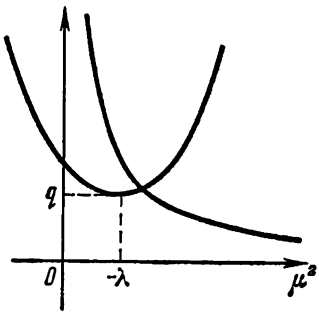


Fig. 184

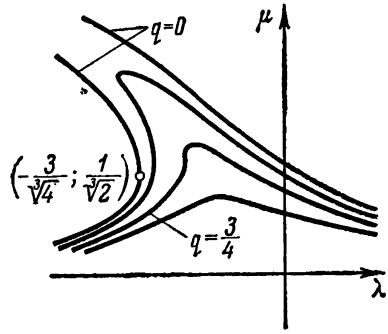


Fig. 185

degree relative to μ^2 there can either be three real roots or one. In Fig. 185 we see the relationship between λ and μ for various fixed values of $q \geqslant 0$. A simple investigation (which we leave to the reader) shows that for $q \geqslant \frac{3}{4}$ the function $\mu(\lambda)$ is one-valued while for $0 \leqslant q < \frac{3}{4}$ there is an interval on which it is three-valued. The above expression of the quantity q indicates that the latter case is observed when the dissipation is relatively small or the external excitation is strong. (The condition $q < \frac{3}{4}$ can be expressed in terms of the dissipation coefficient αh, which makes it possible to verify directly the fulfilment of the condition in numerical examples.) The presence of such an interval $\lambda_2 < \lambda < \lambda_1$ on which $\mu(\lambda)$ is a three-valued function leads to an interesting phenomenon: the amplitude of the forced oscillation changes jump-like when the excitation frequency varies and assumes the values corresponding to $\lambda = \lambda_1$ and $\lambda = \lambda_2$. This phenomenon is completely analogous to the one described in Sec. 3.2 and is illustrated, in the coordinates λ, μ, in Fig. 186. It can be shown that the part

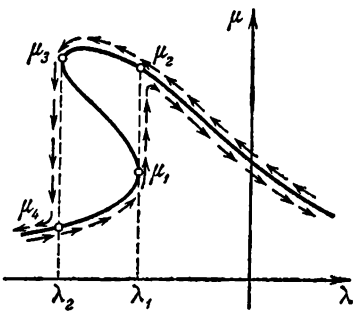


Fig. 186

of the amplitude-frequency characteristic between the points with ordinates μ_1 and μ_2 corresponds to unstable forced vibrations.

Let us discuss the case when the right member of (34) is of the form

$$f_1(x)\alpha + f_2(x)\alpha^2 + \ldots + F_1(t)\alpha + F_2(t)\alpha^2 + \ldots$$

where all the functions $F_j(t)$ are periodic with period T_0. Discarding the terms with F_j's we obtain a nonlinear system for which it is possible to introduce the notion of its natural frequency (see Sec. 2) dependent on the amplitude of oscillations. The value ω_0 may not be in the interval of variation of this natural frequency for $\alpha \neq 0$. In the case of a system with such a right member the

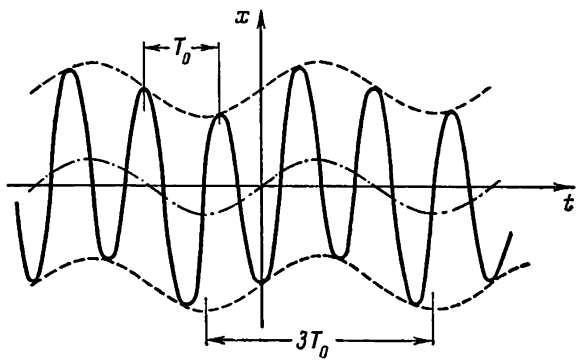


Fig. 187

solution we have constructed above is periodic with period T_0 and thus describes a forced vibration whose frequency coincides with the excitation frequency but this frequency does not necessarily coincide with the natural frequency we have spoken of and is only close to it.

The case when all the functions f_j in (34) have a period mT_0 with respect to t (where m is one of the numbers 2, 3, 4, ...) is treated similarly but possesses a peculiarity mentioned at the end of Sec. 3. Here we use the same formulas (27) and (35). The right-hand member of (36) has the period mT_0 and therefore its Fourier's expansion may involve, in the general case, harmonics with frequencies $\omega_0 = m\left(\frac{2\pi}{mT_0}\right)$ leading to resonance. Consequently, we must impose conditions (37) guaranteeing the absence of such harmonics, find a and b from these determining conditions and then pass to formula (38) with a function $\varphi_1(t)$ of period mT_0 and so on. The resultant solution will have the period mT_0 but its principal term of form (35) will be of period T_0 (see Fig. 187 where $m = 3$). Such

vibrations are said to be **ultraharmonic (superharmonic)**; the frequency of the principal component of such a vibration is an integer multiple of the excitation frequency (and is close to the natural frequency of the unperturbed system). Here again a small external action generates a vibration with finite amplitude, and thus we have a phenomenon which can be interpreted as a resonant vibration whose period is an integer multiple of the natural period. Generally, in nonlinear systems, monofrequent harmonics occur rarely; they usually are "accompanied" with Fourier's expansions involving harmonics with multiple frequencies, that is with fractional periods. In particular, a harmonic with period mT_0 induces a T_0-periodic harmonic which generates resonance.

For systems with several degrees of freedom singular cases are investigated, in principle, with the aid of the same techniques as in the case of one degree of freedom. Of course, formula (35) is then understood in a vector sense and the corresponding constant vectors **a** and **b** include $2k$ arbitrary parameters where k is the number of those natural frequencies of the unperturbed system which are equal to ω_0. In this case the condition guaranteeing the absence of resonant terms reduces to exactly $2k$ scalar equalities, and hence the system of determining equations consists of $2k$ equations in $2k$ unknowns. For $k=1$ this scheme leads to calculations which are not much more complicated than in the case of one degree of freedom.

5. Subharmonic Oscillations. There is an interesting case when the frequency ω of the external excitation is close to $m\omega_0$ where ω_0 is the natural frequency. On the one hand, this case is analogous to the basic case discussed in Sec. 3 since a small external action generates a vibration whose amplitude is also small. On the other hand, as will be shown, together with this small vibration there may also appear a finite one with frequency $\frac{\omega}{m}$ (in Sec. 1 we mentioned that such vibrations were called subharmonic). Hence, here we also have a phenomenon resembling resonance, the excitation frequency being a multiple of the natural frequency. The generation of such vibrations can be explained as follows. If we take a linear equation $\ddot{x}+\omega_0^2 x = A\cos m\omega_0 t$ its general solution is a sum of a particular solution with period $\frac{T_0}{m}$ and the general expression for free vibrations of period T_0 dependent on the initial conditions. Hence, in this simple case we also have a vibration with frequency $\omega_0 = \frac{m\omega_0}{m} = \frac{\omega}{m}$. But if there is an arbitrarily small dissipation in this linear system, free vibrations disappear as $t \longrightarrow \infty$ and only the forced vibration remains. In contrast to it, all the harmonics present in the oscillations of a nonlinear system are inter-

related, and none of them can change without affecting the others. Therefore, when an excitation generates a harmonic with frequency $m\omega_0$ the latter may induce a harmonic of frequency ω_0 which results in a resonant vibration.

This phenomenon can be described mathematically in the following way. Suppose that all the functions f_j in the right-hand side of (34) have a period $\frac{T_0}{m}$ with respect to t. Then they also have T_0 as their period, and hence we can pass to formula (35) and determining equations (37). In this case the values $a=b=0$ always

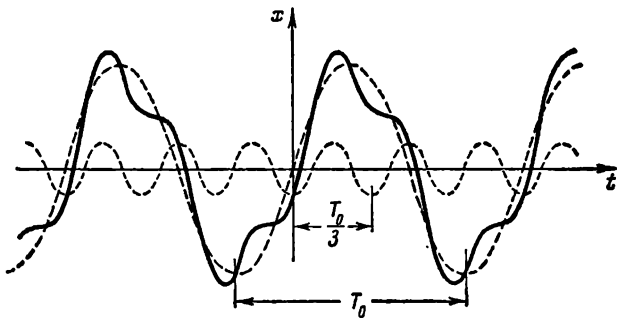


Fig. 188

satisfy these equations and lead to the solution constructed in Sec. 3. But equations (37) may also possess nonzero solutions which thus generate subharmonic vibrations. A possible shape of such vibrations is depicted in Fig. 188 where $m=3$.

As an instance, let us take one more modification of Duffing's equation:

$$\ddot{x}+\omega_0^2 x=\alpha p x+\alpha k\,(x+A\cos 3\omega_0 t)^3$$

Determining equations (37) in this particular case are written, after the integration has been performed, in the form (check it up!)

$$\frac{4}{3}\frac{p}{k}a+a^3+ab^2+a^2A+2aA^2-b^2A=0,$$

$$b\left(\frac{4}{3}\frac{p}{k}+b^2+a^2+2A^2-2aA\right)=0$$

We see that besides the trivial solution $a=b=0$ these determining equations possess nonzero solutions, namely

$$a=-\frac{A}{2}\pm\left(-\frac{7}{4}A^2-\frac{4}{3}\frac{p}{k}\right)^{\frac{1}{2}},\quad b=0$$

(obtained from the equation $a^2+Aa+2A^2+\frac{4}{3}\frac{p}{k}=0$ into which

The first determining equation turns for $b=0$ and

$$a=\frac{A}{4}\pm\sqrt{-\frac{3}{16}A^2-\frac{1}{3}\frac{p}{k}}, \quad b=\pm\sqrt{3a}$$

(determined from the equation $4a^2-2Aa-2A^2+\frac{4}{3}\frac{p}{k}=0$ resulting from the substitution of $b=\pm\sqrt{3a}$ into the second determining equation). These solutions are valid only when the given combinations of the parameters do not yield imaginary roots of the determining equations.

6. Notes on Forced Oscillations. Here we give some additional remarks concerning forced oscillations.

1) In the investigation of forced oscillations the so-called method of harmonic balance is widely used whose feature is that, in principle, it is applicable to equations not involving a small parameter. For example, suppose that the left-hand side of the equation

$$f(x,\ \dot{x},\ \ddot{x},\ t)=0 \tag{43}$$

is periodic in t with period T and that we are interested in a solution of the equation having the same period. (If the periods of the functions $f(x,\ \dot{x},\ \ddot{x},\ t)$ and of the desired solution do not coincide but are commensurable, that is are integral multiples of a common period, the latter can be taken as T.) The sought-for solution can be expanded into Fourier's series

$$x(t)=a_0+\sum_{j=1}^{\infty}(a_j\cos j\omega t+b_j\sin j\omega t), \quad \omega=\frac{2\pi}{T}, \tag{44}$$

with undetermined coefficients. The condition that this solution satisfies equation (43) can be written in the form of the two sequences of equalities

$$\left.\begin{aligned}&\int_0^T f\Big(a_0+\sum_{j=1}^{\infty}(a_j\cos j\omega t+b_j\sin j\omega t),\ \omega\sum_{j=1}^{\infty}j(-a_j\sin j\omega t+b_j\cos j\omega t),\\ &\qquad -\omega^2\sum_{j=1}^{\infty}j^2(a_j\cos j\omega t+b_j\sin j\omega t),\ t\Big)\cos k\omega t\,dt=0,\\ &\qquad k=0,\ 1,\ 2,\ \ldots\\ &\int_0^T f\Big(a_0+\sum_{j=1}^{\infty}(a_j\cos j\omega t+b_j\sin j\omega t),\ \omega\sum_{j=1}^{\infty}j(-a_j\sin j\omega t+b_j\cos j\omega t),\\ &\qquad -\omega^2\sum_{j=1}^{\infty}j^2(a_j\cos j\omega t+b_j\sin j\omega t),\ t\Big)\sin k\omega t\,dt=0,\\ &\qquad k=1,\ 2,\ 3\ldots\end{aligned}\right\} \tag{45}$$

Thus, for the coefficients of expansion (44), we have obtained a system of an infinite number of nonlinear equations with an infinite number of unknowns. To solve the system we can truncate it, that is take equations (45) for $k = 1, 2, \ldots, k_0$ and extend the summation in (44) and in all the other expansions involved only over the values $j = 1, 2, \ldots, k_0$ of the index j where k_0 is a chosen fixed number. This means that the discrepancy arising from the substitution of this finite sum of k_0 harmonics into equation (43) is subject to the requirement (the "balance" condition) that its Fourier's expansion does not contain the harmonics entering into the sum representing the approximation to the sought-for solution. Thus, for $k_0 = 0$ we obtain one equation

$$\int_0^T f(a_0, 0, 0, t)\, dt = 0$$

in one unknown, for $k_0 = 1$ the corresponding system of three equations with three unknowns, etc. Passing from k_0 to the subsequent value $k_0 + 1$ we can apply an iteration scheme taking as a zeroth approximation the solution found for k_0. If we have some additional information about the properties of the solution, for instance, if it is known that $x(t)$ is an even or an odd function or that there are certain dominant harmonics in its Fourier's expansion, this information should be taken into account when the balance conditions are introduced since this leads to a substantial increase of the accuracy of the approximate solution for the same amount of calculations.

2. Calculations may show that for the given values of the parameters of a system and the given excitation (external action) there can be several stationary vibrations (that is periodic solutions), and then the question arises which of them is in fact realized. In real systems only stable solutions can naturally be realized, and therefore this question reduces to an investigation of the stability of the solutions. The latter problem can be solved by means of studying the linearized equation (see Sec. 3.4) which is a linear equation with periodic coefficients. In particular, for a system with one degree of freedom, that is for a scalar equation of the second order, the application of the transformations indicated in Sec. 1.10 leads to Hill's equation (Sec. 1.3). If there are several stable stationary regimes of vibration each of them has its own region of initial data within which the neighbouring solutions are attracted to the given solution. When the system is subjected to perturbations the solution may change in a jump-like manner and pass to another stable regime. At present there are no general analytic methods for determining these regions but it is sometimes

possible to find them by means of numerical integration on digital or analogue computers for chosen initial data.

3. In conclusion, let us discuss the case when the external excitation of a nonlinear system is described by a **polyharmonic** (that is quasi-periodic) function with a finite number of basic frequencies which, in the general case, are incommensurable (see Sec. 1.7). For definiteness, take the equation

$$\ddot{x} + \omega_0^2 x = \alpha \sum_{j=0}^{\infty} a_j x^j + A_1 \cos \omega_1 t + A_2 \cos \omega_2 t \qquad (46)$$

where for each of the frequencies ω_1 and ω_2 the basic case studied in Sec. 3 takes place. As in the foregoing sections,

$$x_0 = \frac{A_1}{\omega_0^2 - \omega_1^2} \cos \omega_1 t + \frac{A_2}{\omega_0^2 - \omega_2^2} \cos \omega_2 t$$

which leads to the equation

$$\ddot{x}_1 + \omega_0^2 x_1 = \sum_{j=0}^{\infty} a_j \left(\frac{A_1}{\omega_0^2 - \omega_1^2} \cos \omega_1 t + \frac{A_2}{\omega_0^2 - \omega_2^2} \cos \omega_2 t \right)^j$$

for $x_1(t)$. To solve this equation we must open the brackets on the right-hand side and transform each of the resulting terms into a sum of cosines. It is clear that this results in terms of the form $A \cos (m_1 \omega_1 + m_2 \omega_2) t$ with arbitrary integers m_1 and m_2 of any sign. Therefore, the expression for x_1 will contain terms of the form

$$\frac{A}{\omega_0^2 - (m_1 \omega_1 + m_2 \omega_2)^2} \cos (m_1 \omega_1 + m_2 \omega_2) t$$

If ω_1 and ω_2 are incommensurable, that is the quotient $\frac{\omega_1}{\omega_2}$ is an irrational number, an expression of the form $\omega_0 - (m_1 \omega_1 + m_2 \omega_2) = \omega_2 \left\{ \frac{\omega_0}{\omega_2} - \left[m_1 \frac{\omega_1}{\omega_2} + m_2 \right] \right\}$ taken for any given ω can be made arbitrarily small for some appropriately chosen m_1 and m_2. (This immediately follows from the assertion, mentioned in Sec. 2.13, that for incommensurable frequencies the trajectories on the torus are everywhere dense.) As was pointed out by H. Poincaré, the presence of these "**small divisors**" leads to difficulties in proving the convergence of the approximations of the small-parameter method. If the right-hand side of (46) contains a finite sum instead of the infinite series the expressions of x_j's are also finite sums but the same difficulty appears again when we pass to the limit as $j \to \infty$. In calculations connected with real problems this difficulty is sometimes overcome, when necessary, by passing to smaller values of α.

7. Auto-Oscillations. We shall investigate the differential equation

$$\ddot{x}+\omega_0^2 x=\alpha f_1(x,\ \dot{x})+\alpha^2 f_2(x,\ \dot{x})+\ldots \tag{47}$$

containing a small parameter α and describing an autonomous system with one degree of freedom. In Sec. 2.5 we mentioned that such a system may have limit cycles corresponding to auto-oscillations. Here we shall show how they can be determined.

An essential distinction between forced oscillations and auto-oscillations is that the frequency ω of an auto-oscillation is not set beforehand but is found in the calculation process. Therefore, we take the expansion

$$\omega=\omega_0+\omega_1\alpha+\omega_2\alpha^2+\ldots \tag{48}$$

in which all the coefficients except ω_0 are undetermined and make the change $\omega t=\theta$ of the independent variable. Then equation (47) is rewritten in the form

$$\begin{aligned}(\omega_0+\omega_1\alpha+\omega_2\alpha^2+\ldots)^2\ddot{x}+\omega_0^2 x=\\ =\alpha f_1(x,\ (\omega_0+\omega_1\alpha+\omega_2\alpha^2+\ldots)\dot{x})+\ldots\end{aligned} \tag{49}$$

where the dot is now used for designating differentiation with respect to the new independent variable θ. As before, the solution is looked for in the form of an expansion in powers of α:

$$x=x_0(\theta)+x_1(\theta)\alpha+x_2(\theta)\alpha^2+\ldots \tag{50}$$

The substitution of (50) into (49) leads to the relations

$$\left.\begin{aligned}&\ddot{x}_0+x_0=0\\ &\ddot{x}_1+x_1=-\frac{2\omega_1}{\omega_0}\ddot{x}_0(\theta)+\frac{1}{\omega_0^2}f_1(x_0(\theta),\ \omega_0\dot{x}_0(\theta))\\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\end{aligned}\right\} \tag{51}$$

Another peculiarity of autonomous vibrations is that the initial phase can be chosen in an arbitrary way. This means that any point of the cycle in question can be taken as an initial point corresponding to time moment $t=0$ (this is not the case for forced oscillations). In particular, we can put $x(0)=0$

$$x_0(0)=0,\ x_1(0)=0,\ x_2(0)=0,\ \ldots \tag{52}$$

Now we can proceed to determine, in succession, the solutions of equations (51). From the first equation (51), taking into account conditions (52), we obtain the expression

$$x_0=b\sin\theta \tag{53}$$

where the amplitude b is yet unknown. Substituting (53) into the right-hand side of the second equation (51) and imposing the

nonresonance condition we obtain the equalities

$$\int_0^{2\pi} [2\omega_0\omega_1 b \sin\theta + f_1(b\sin\theta, \omega_0 b\cos\theta)]\cos\theta\, d\theta = 0$$

and

$$\int_0^{2\pi} [2\omega_0\omega_1 b \sin\theta + f_1(b\sin\theta, \omega_0 b\cos\theta)]\sin\theta\, d\theta = 0$$

which can be rewritten in the form

$$\left.\begin{aligned} &\int_0^{2\pi} f_1(b\sin\theta, \omega_0 b\cos\theta)\cos\theta\, d\theta = 0 \\ &\omega_1 = -\frac{1}{2\pi b\omega_0}\int_0^{2\pi} f_1(b\sin\theta, \omega_0 b\cos\theta)\sin\theta\, d\theta \end{aligned}\right\} \qquad (54)$$

The first equality (54) is a determining equation for the amplitude b. After b has been found the second equality (54) yields an explicit expression for ω_1. Then we take the second equation (51) and find from it the expression $x_1 = \varphi_1(\theta) + b_1\sin\theta\,(\varphi_1(0) = 0)$; the condition of the absence of resonant oscillations for the subsequent equation results in two determining equations from which b_1 and ω_2 are found and so on. (Let the reader carry out an analogous investigation by making in equation (47), as in Sec. 4, the change $\omega_0 = \omega - \omega_1\alpha - \ldots$, and explain the difference between the results.)

To test the cycle thus constructed for stability we must write (47) in the form of a system of first-order equations and then apply the general condition (2.22). The computations which we leave to the reader show that if

$$\alpha\int_0^{2\pi} f'_{\dot{x}}(b\sin\theta, \omega_0 b\cos\theta)\, d\theta < 0 \qquad (55)$$

the cycle is stable and if the left member of (55) is positive it is unstable.

Let us apply this construction to the **van der Pol equation**

$$\ddot{x} + \omega_0^2 x = \alpha(\dot{x} - k\dot{x}^3) \quad (k = \text{const} > 0) \qquad (56)$$

For $\alpha > 0$ and small values of the velocity $\dot{x}$ this equation looks as if it describes an oscillator with negative damping (why?); this indicates that the origin in the phase plane is an unstable focal point. For large values of the velocity $\dot{x}$ the term $-k\dot{x}^3$ dominates over $\dot{x}$, that is the solution is rapidly damped. Hence, it is

natural to expect that for "intermediate" values of $\dot{x}$ there exists a stable limit cycle. Let the reader show that in this case the determining equation has the form

$$\pi\omega_0 b - \frac{3}{4}\pi k\omega_0^3 b^3 = 0$$

whence $b = \pm 2(\omega_0\sqrt{3k})^{-1}$. The solutions of equation (56) being invariant with respect to the transformation $x \to -x$, both values of b yield the same cycle. Therefore, we can take one of them, say $b = -2(\omega_0\sqrt{3k})^{-1}$. Now applying criterion (55) we conclude that the constructed cycle is stable for $\alpha > 0$, which, by the way, obviously follows from the argument at the beginning of this paragraph. The second equation (54) gives us the value $\omega_1 = 0$; this means that to analyse the variation of the frequency it is necessary to use the subsequent terms of the expansion. (The equality $\omega_1 = 0$ is a direct consequence of the fact that equation (56) goes into itself under the transformation $\alpha \to -\alpha$, $t \to -t$ because this implies that ω is an even function of the parameter α, and therefore its expansion in powers of α can only involve even powers.) The computations yield

$$x_1 = \frac{2}{3\omega_0^2\sqrt{3k}}(\cos\theta - \cos 3\theta) + b_1 \sin\theta$$

and

$$\omega_0^2(\ddot{x}_2 + x_2) = -2\omega_0\omega_2\ddot{x}_0 + \omega_0\dot{x}_1 - 3k\omega_0^3\dot{x}_0^2\dot{x}_1 =$$

$$= \frac{4}{\sqrt{3k}}\omega_2 \sin\theta + \frac{2}{3\omega_0\sqrt{3k}}(-\sin\theta + 3\sin 3\theta) +$$

$$+ \omega_0 b_1 \cos\theta - 4\omega_0\cos^2\theta\left[\frac{2}{3\omega_0^2\sqrt{3k}}(-\sin\theta + 3\sin 3\theta) + b_1\cos\theta\right]$$

Applying the condition guaranteeing the absence of resonant terms in $x_2(t)$ we find (check it up!) $b_1 = 0$ and $\omega_2 = \frac{1}{2\omega_0}$.

Let the reader verify that the application of the method of Sec. 4 to equation (56) only leads to the zero solution.

In conclusion we remark that auto-oscillations most often occur when the autonomous system in question absorbs the energy of an external energy source sustaining the oscillations, and there is dissipation in the system which prevents the amplitude of the oscillations from becoming too large. Such systems are referred to as active ones in contrast to passive systems which have no external energy sources and whose total energy either retains a constant value or decreases. The well-known examples of such self-sustained oscillations are the vibration of a violin string excited by the bow, the creaking of the door hinges, the vibration of

telegraph wires produced by the wind, the flatter of aeroplane wings, etc.

The above method can also be applied to investigating small vibrations of a strongly nonlinear conservative system described by an equation of the form $\ddot{x}+\omega_0^2 x=f(x)$ where $f(0)=0$. In this problem the role of the small parameter is played by the amplitude of oscillations. For instance, we can put $x(0)=\alpha$, $\dot{x}(0)=0$. Here we must also use expansion (48) and the change $\omega t=\theta$ but expression (50) should be replaced by $x=x_1(\theta)\alpha+x_2(\theta)\alpha^2+\ldots$ where $x_1(0)=1$, $x_2(0)=x_3(0)=\ldots=0$, $\dot{x}_1(0)=\dot{x}_2(0)=\ldots=0$. Let the reader carry out the calculations.

8. **Relaxation Oscillations.** Let us take equation (47) of a more special type

$$\ddot{x}+\omega_0^2 x=\alpha f(\dot{x})$$

involving a large parameter α. We thus are interested in the behaviour of the solution for $\alpha\to\infty$. The introduction of the new variables $\dot{x}=y$, $x=\alpha\xi$, $t=\alpha\tau$ and $\alpha^{-2}=\varepsilon$ leads to the system of equations

$$\frac{d\xi}{d\tau}=y,\quad \varepsilon\frac{dy}{d\tau}=-\omega_0^2\xi+f(y) \tag{57}$$

(check up the calculations!) where ε is small.

System (57) is a particular case of a more general system

$$\frac{dx}{d\tau}=P(x, y, \varepsilon),\quad \varepsilon\frac{dy}{d\tau}=Q(x, y, \varepsilon) \tag{58}$$

with small parameter $\varepsilon>0$. The second equation can be written in the form $\frac{dy}{d\tau}=\frac{1}{\varepsilon}Q(x, \dot{y}, \varepsilon)$, which shows that when ε is small the behaviour of the trajectory in a chosen part of the x y-plane is essentially specified by the sign of the function $Q(x, y, 0)$ for the values of x and y corresponding to this part. For instance, if $Q(x, y, 0)>0$ the vertical component of the velocity of the flow is directed upward and is large for small values of ε while the horizontal component is finite; this means that the flow is directed almost vertically upward.

The line (L) determined by the equation $Q_0(x, y)\equiv Q(x, y, 0)=0$ divides the x y-plane into parts in which the flow is directed almost vertically upward or downward (see Fig. 189). Near the line (L) the velocity of the flow remains finite, and therefore in the vicinity of (L) the moving point may go along (L). The direction of this motion is specified by the sign of the function $P(x, y, 0)$ on (L). It is also necessary to distinguish between stable arcs of (L) (such as the arcs AB and CD in Fig. 189) and unstable ones (such as the arc BC in the figure). (There can also be half-

stable arcs which will not be considered here.) The trajectories are attracted to the stable arcs: when a trajectory falls on such an arc or is near it, the moving point goes along that stable arc until it passes to an unstable arc (in Fig. 189 such a situation is shown near the point B). A trajectory going along an unstable arc of the line (L) has a tendency to break away from it, and therefore such a motion is not realizable practically.

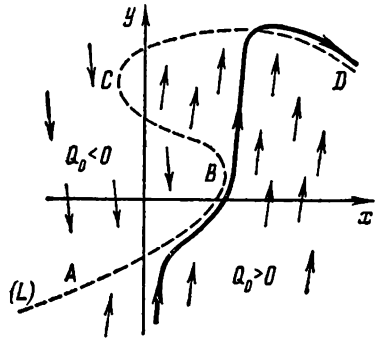


Fig. 189

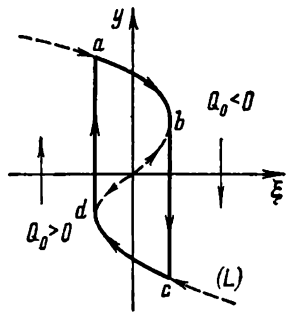


Fig. 190

The sign of the function $P(x, y, 0)$ may change along (L) in such a way that the trajectory periodically describes a path approaching a limit cycle which, for $\varepsilon \to 0$, consists of a number of stable arcs of (L) traversed with a finite velocity and of several vertical segments which are passed with an infinitely large velocity. (Here of course we speak of the velocity of the representing point, in the phase plane, related to unit measure of τ.) Consequently, for small nonzero values of ε this vibration consists of slow and rapid stages which usually correspond, respectively, to processes of slow accumulation of energy and its rapid losses. An oscillation of this type is referred to as a **relaxation vibration.**

Let us come back to system (57) and suppose that the graph (L) of the function $\xi = \frac{1}{\omega_0^2} f(y)$ is of the form shown in Fig. 190. Then, for $\varepsilon \to 0$ the system has, in the limit, the cycle (C) shown in Fig. 190 by the continuous line. The corresponding period of motion is readily found: we have $d\tau = \frac{d\xi}{y}$ and therefore

$$T = \int\limits_{ab} \frac{d\xi}{y} + \int\limits_{cd} \frac{d\xi}{y} = \frac{1}{\omega_0^2}\left(\int\limits_{y_a}^{y_b} \frac{f'(y)}{y}\, dy + \int\limits_{y_c}^{y_d} \frac{f'(y)}{y}\, dy \right)$$

(The same technique can be extended to the general system (58).) In particular, in equation (56) we have $f(y) = y - ky^3$, and thus

$y_a=\frac{2}{\sqrt{3k}}$ and $y_b=\frac{1}{\sqrt{3k}}$ (check it up!). Substituting these values into the above expression we finally find the period expressed in units of τ:

$$T=\frac{1}{\omega_0^2}\,2\int\limits_{\frac{2}{\sqrt{3k}}}^{\frac{1}{\sqrt{3k}}}\frac{1-3ky^2}{y}\,dy=(3-2\ln 2)\,\omega_0^{-2}=1.6137\omega_0^{-2}$$

If α is large but finite the limit cycle is close to the one depicted in Fig. 190. To return to real time t from the "slow time" τ we must multiply the above expression by α, which results in the principal term of the asymptotic expansion of the period of vibrations. The Soviet scientist A.A. Dorodnitsyn (born in 1910) obtained in 1947 subsequent terms of this expansion (see [12]):

$$T=\frac{1}{\omega_0}\left[1.6137\frac{\alpha}{\omega_0}+7.014\sqrt[3]{\frac{\omega_0}{\alpha}}-\frac{22}{9}\frac{\omega_0}{\alpha}\ln\frac{\alpha}{\omega_0}+\right.$$
$$\left.+0.0087\frac{\omega_0}{\alpha}+O\left(\left(\frac{\omega_0}{\alpha}\right)^{\frac{4}{3}}\right)\right]$$

A jump-like passage from one type of motion to another is also found in free vibrations of a linear oscillator with small inertia term described by the equation

$$m\ddot{x}+f\dot{x}+kx=0 \tag{59}$$

with constant coefficients and a very small value of m. In the limit, making m tend to zero, we obtain the first-order equation $f\dot{x}+kx=0$ (therefore in this case we sometimes speak, conditionally, about "$\frac{1}{2}$ degree of freedom"). The system of equations describing the phase trajectories of equation (59) is written as

$$\dot{x}=y,\quad m\dot{y}=-fy-kx$$

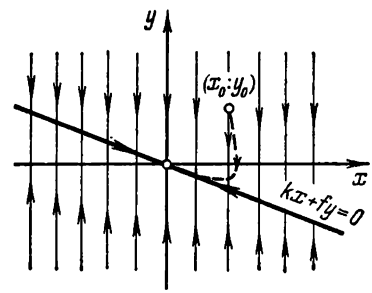


Fig. 191

The phase diagram for infinitesimal values of m is shown in Fig. 191. For any initial point (x, y) in the phase plane not lying on the straight line $kx+fy=0$ the moving point instantaneously jumps to this straight line (this is possible because of the smallness of the measure m of inertia). After that the point representing the state of the oscillator passes with a finite speed, according to an

exponential law, to the equilibrium position $x=y=0$. When m is small but finite (not infinitesimal) the trajectory in the phase plane has, of course, the shape shown in Fig. 191 by the dotted line.

9. Boundary Layer. In connection with the subject matter of the foregoing section we shall dwell in more detail on systems of differential equations involving a small parameter $\alpha > 0$ which become completely or partially degenerate for $\alpha = 0$. Consider a system of the form

$$\dot{\mathbf{x}} = \mathbf{f}(t, x, y, \alpha), \quad \alpha\dot{\mathbf{y}} = \mathbf{g}(t, x, y, \alpha) \tag{60}$$

where the sought-for vectors $\mathbf{x}(t)$ and $\mathbf{y}(t)$ are, respectively, of dimensions n and m. For $\alpha = 0$ the second group of equations (60) degenerates into the system of finite equations

$$\mathbf{g}(t, \mathbf{x}, \mathbf{y}, 0) = \mathbf{0} \tag{61}$$

($x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_m)$) specifying an $(n+1)$-dimensional hypersurface (S) in the $(n+m+1)$-dimensional space of the variables t, x, y.

For small values of α every integral curve $x(t)$, $y(t)$ of system (60) may have, as in Sec. 8, parts of the two following types: those lying at a finite distance from (S) and those passing close to (S). On the parts of the first type we can consider, to within infinitesimals of the order of α, the quantities t and x as being constant and y as a solution of the system of equations

$$\frac{dy}{d\tau} = \mathbf{g}(t, x, y, 0) \tag{62}$$

where $\tau \left(d\tau = \frac{dt}{\alpha} \right)$ is the "**fast time**". For the parts of the second type the system

$$\frac{dx}{dt} = \mathbf{f}(t, x, y, 0) \tag{63}$$

(where $y = \chi(t, x)$ is found from system (61)) is approximately satisfied.

Suppose that we are given initial conditions

$$x|_{t=0} = x_0, \quad y|_{t=0} = y_0$$

specifying a point M_0 in E_{n+m+1} which, in the general case, does not lie on the surface (S). Then the initial portion of the integral curve issued from M_0 is a part of the first type and is determined by the autonomous system (62) taken with $t=0$ and $x=x_0$. We shall limit ourselves to the case when the point y_0 belongs to the domain where the trajectories of system (62) are attracted to an asymptotically stable rest point $\tilde{y}_0$ of this system. Therefore, we shall assume that $(0, x_0, \tilde{y}_0) \in (S)$ and that all the eigenvalues of

the matrix $\left.\frac{\partial \mathbf{g}(0, x_0, y, 0)}{\partial y}\right|_{y=\tilde{y}_0}$ have negative real parts. The last condition implies, in particular, that $\det\left(\frac{\partial g}{\partial y}\right)\Big|_{(0, x_0, \tilde{y}_0)} \neq 0$, which indicates that in the vicinity of the point $\tilde{M}_0(0, x_0, \tilde{y}_0)$ the equation of the surface (S) can be taken in the form $y = \chi(t, x)$, which makes it possible to consider equation (63) on the surface (S).

Thus, for small values of α the integral curve under consideration consists of a part along which the moving point rapidly passes from M_0 to a small neighbourhood of the point $\tilde{M}_0$ and of a part along which the moving point, after it has arrived at the neighbourhood of $\tilde{M}_0$, travels in the vicinity of (S) in accordance with the degenerate equation (63). The interval of variation of the independent variable t corresponding to the first part is called a **boundary layer**. On this interval the solution passes from an arbitrarily chosen initial point $(0, x_0, y_0)$ to the point $(0, x_0, \tilde{y}_0)$ whose coordinates satisfy the degenerate system (61). The width of the boundary layer is understood conditionally and is usually introduced in such a way that, to within infinitesimals of higher order, it is inversely proportional to the slope of the field specified by system (60), that is directly proportional to α (e. g. cf. [107], Sec. XV.28).

It should also be mentioned that an integral curve going along (S) may break away from (S) only when the above condition of asymptotic stability of the rest point is violated.

To derive a more accurate asymptotic representation of the solution we are interested in we must take into account that, according to the above-described behaviour of the solution, its asymptotic expansion must contain both functions of τ playing a dominant role in the boundary layer and functions of t dominant for $t \gg \alpha$. Therefore, we write the expansion in the form

$$\left.\begin{aligned} \mathbf{x} &= \boldsymbol{\varphi}_0(t) + \alpha\boldsymbol{\varphi}_1(t) + \ldots + \boldsymbol{\Phi}_0(\tau) + \alpha\boldsymbol{\Phi}_1(\tau) + \ldots \\ \mathbf{y} &= \boldsymbol{\psi}_0(t) + \alpha\boldsymbol{\psi}_1(t) + \ldots + \boldsymbol{\Psi}_0(\tau) + \alpha\boldsymbol{\Psi}_1(\tau) + \ldots \end{aligned}\right\} \tag{64}$$

A representation of this type is of course not uniquely determined; for instance, if $\boldsymbol{\varphi}_0(t) \equiv t$ and $\boldsymbol{\Phi}_1(\tau) \equiv 0$ we can equivalently write $\boldsymbol{\varphi}_0(t) \equiv 0$ and $\boldsymbol{\Phi}_1(\tau) \equiv \tau$. But this ambiguity is inessential for the construction presented below.

Let us substitute expansion (64) into (60) and then transform the resultant equations according to the following scheme:

$$\begin{aligned} &\boldsymbol{\varphi}_0'(t) + \alpha\boldsymbol{\varphi}_1'(t) + \ldots + \frac{1}{\alpha}\left[\boldsymbol{\Phi}_0'(\tau) + \alpha\boldsymbol{\Phi}_1'(\tau) + \ldots\right] = \\ &= \mathbf{f}(t, \boldsymbol{\varphi}_0(t) + \alpha\boldsymbol{\varphi}_1(t) + \ldots, \boldsymbol{\psi}_0(t) + \alpha\boldsymbol{\psi}_1(t) + \ldots, \alpha) + \\ &+ [\mathbf{f}(\alpha\tau, \boldsymbol{\varphi}_0(\alpha\tau) + \alpha\boldsymbol{\varphi}_1(\alpha\tau) + \ldots + \boldsymbol{\Phi}_0(\tau) + \alpha\boldsymbol{\Phi}_1(\tau) + \ldots, \\ &\boldsymbol{\psi}_0(\alpha\tau) + \alpha\boldsymbol{\psi}_1(\alpha\tau) + \ldots + \boldsymbol{\Psi}_0(\tau) + \alpha\boldsymbol{\Psi}_1(\tau) + \ldots, \alpha) - \mathbf{f}(\alpha\tau, \boldsymbol{\varphi}_0(\alpha\tau) + \\ &+ \alpha\boldsymbol{\varphi}_1(\alpha\tau) + \ldots, \boldsymbol{\psi}_0(\alpha\tau) + \alpha\boldsymbol{\psi}_1(\alpha\tau) + \ldots, \alpha)] \end{aligned}$$

(for the second equation (60) the transformation is carried out in the same manner). Next we expand the right-hand sides in powers of α and equate the coefficients in like powers of α separately for functions of t and for functions of τ. This first of all yields $\Phi_0'(\tau) \equiv 0$ whence it follows that we can put $\Phi_0(\tau) \equiv 0$ since the constant term can be attributed to φ_0. Furthermore, we obtain

$$\varphi_0'(t) = \mathbf{f}(t, \varphi_0(t), \psi_0(t), 0), \quad \mathbf{0} = \mathbf{g}(t, \varphi_0(t), \psi_0(t), 0) \tag{65}$$

and

$$\Psi_0'(\tau) = \mathbf{g}(0, \varphi_0(0), \psi_0(0) + \Psi_0(\tau), 0) - \mathbf{g}(0, \varphi_0(0), \psi_0(0), 0) \tag{66}$$

Besides, the initial conditions

$$\varphi_0(0) + \Phi_0(0) = \mathbf{x}_0, \quad \psi_0(0) + \Psi_0(0) = \mathbf{y}_0 \tag{67}$$

must be fulfilled. From (65) and (67) we see that the functions $x = \varphi_0(t)$, $y = \psi_0(t)$ satisfy the limiting system, obtained from (60) for $\alpha = 0$, with the initial conditions $\varphi_0(0) = x_0$, $\psi_0(0) = \tilde{y}_0$. According to (66), this enables us to rewrite the initial-value problem for the principal term $\Psi_0(\tau)$, corresponding to the boundary layer, in the form

$$\Psi_0'(\tau) = \mathbf{g}(0, \varphi_0(0), \psi_0(0) + \Psi_0(\tau), 0), \quad \Psi_0(0) = \mathbf{y}_0 - \tilde{\mathbf{y}}_0$$

Thus, the function $\mathbf{y} = \tilde{\mathbf{y}}_0 + \Psi_0(\tau)$ satisfies system (62) for $t = 0$, $\mathbf{x} = \mathbf{x}_0$. Besides, the hypothesis on the eigenvalues of the matrix $\frac{\partial \mathbf{g}}{\partial \mathbf{y}}$ implies that $\Psi_0(\tau)$ tends to zero at an exponential rate as $\tau \to \infty$.

Equating subsequent terms of the expansion we receive the equations

$$\varphi_1'(t) = \mathbf{f}_x'(t, \varphi_0(t), \psi_0(t), 0)\varphi_1(t) + \mathbf{f}_y'(t, \varphi_0(t), \psi_0(t), 0)\psi_1(t) + \mathbf{f}_\alpha'(t, \varphi_0(t), \psi_0(t), 0) \tag{68}$$

$$\psi_0'(t) = \mathbf{g}_x'(t, \varphi_0(t), \psi_0(t), 0)\varphi_1(t) + \mathbf{g}_y'(t, \varphi_0(t), \psi_0(t), 0)\psi_1(t) + \mathbf{g}_\alpha'(t, \varphi_0(t), \psi_0(t), 0) \tag{69}$$

$$\Phi_1'(\tau) = \mathbf{f}(0, \varphi_0(0), \psi_0(0) + \Psi_0(\tau), 0) - \mathbf{f}(0, \varphi_0(0), \psi_0(0), 0) \tag{70}$$

$$\begin{aligned}\Psi_1'(\tau) = {} & \{\mathbf{g}_t'(0, \varphi_0(0), \psi_0(0) + \Psi_0(\tau), 0)\tau + \mathbf{g}_x'(0, \varphi_0(0), \psi_0(0) + \\ & + \Psi_0(\tau), 0)[\varphi_0'(0)\tau + \varphi_1(0) + \Phi_1(\tau)] + \mathbf{g}_y'(0, \varphi_0(0), \psi_0(0) + \\ & + \Psi_0(\tau), 0)[\psi_0'(0)\tau + \psi_1(0) + \Psi_1(\tau)] + \mathbf{g}_\alpha'(0, \varphi_0(0), \psi_0(0) + \\ & + \Psi_0(\tau), 0)\} - \mathbf{g}_t'(0, \varphi_0(0), \psi_0(0), 0)\tau + \\ & + \mathbf{g}_x'(0, \varphi_0(0)\psi_0(0), 0)[\varphi_0'(0)\tau + \varphi_1(0)] + \\ & + \mathbf{g}_y'(0, \varphi_0(0), \psi_0(0), 0)[\psi_0'(0)\tau + \psi_1(0)] + \\ & + \mathbf{g}_\alpha'(0, \varphi_0(0), \psi_0(0), 0)\}\end{aligned} \tag{71}$$

with the initial conditions

$$\varphi_1(0) + \Phi_1(0) = \mathbf{0}, \quad \psi_1(0) + \Psi_1(0) = \mathbf{0} \tag{72}$$

From equation (70) and the condition $\Phi_1(\infty) = 0$ we find

$$\Phi_1(\tau) = \int_{\tau}^{\infty} [f(0,\ \varphi_0(0),\ \psi_0(0),\ 0) - f(0,\ \varphi_0(0),\ \psi_0(0) + \Psi_0(\tau_1), 0)]\, d\tau_1$$

whence, in particular, $\Phi_1(0)$ is determined, and therefore, by virtue of (72), $\varphi_1(0)$ is also found. Expressing the unknown function $\psi_1(t)$ from (69) in terms of $\varphi_1(t)$ and substituting the result into (68) we arrive at a system of the first order with respect to φ_1 from which, using the known value $\varphi_1(0)$, we find $\varphi_1(t)$ and hence $\psi_1(t)$ as well. Next we can determine $\Psi_1(0)$ by means of (72) and the function $\Psi_1(\tau)$ by means of (71). It can also be easily shown (prove it!) that for $\tau \to \infty$ the functions $\Phi_1(\tau)$ and $\Psi_1(\tau)$ tend to zero at an exponential rate.

The subsequent terms of expansion (64) are found in a similar way. A more thorough analysis shows that the resultant series are asymptotically convergent for $\alpha \to 0$ to the solution of the initial-value problem for every interval $0 \leqslant t \leqslant T$ on which the solution of the limiting (degenerate) problem goes along the part of the surface (S) for which the above condition of the asymptotic stability holds.

10. Nonperiodic Oscillations. The methods described above are mainly applicable to investigating and constructing solutions describing periodic oscillations (both forced and free ones). For transient processes in vibration systems, almost periodic and other nonperiodic types of vibration it is convenient to use a more general method inaugurated in 1927 by van der Pol and extensively developed by A.A. Andronov (in 1930), N.M. Krylov and N.N. Bogolyubov (in 1937) and later by Yu.A. Mitropolsky and other scientists (see [13], [91] and [104]). Here we shall demonstrate some of these techniques.

The method described in Sec. 3 is not only applicable to constructing periodic solutions but can also be modified for investigating transient processes. In this connection let us consider the singular case for equation (34) with periodic functions f_j of period T_0 $\left(T_0 = \frac{2\pi}{\omega_0}\right)$ with respect to t. As was indicated in Sec. 3, to this problem the investigation of the excitation of a system with small nonlinearity is reduced when the excitation frequency is close to the resonance frequency.

We shall consider the "ordinary" time t and also the "**slow time**" $\tau = \alpha t$. The basic idea of the method is to construct the solutions in the form

$$x = \sum_{k=0}^{\infty} a_k(t,\ \tau)\, \alpha^k \tag{73}$$

(this series may be only asymptotically convergent for $\alpha \to 0$; see Sec. II.4.1). Such a representation of the solution is of course not uniquely determined; for instance, this is readily seen from the simplest relation $\tau + 0 \cdot \alpha = 0 + t\alpha$. But this ambiguity makes it possible to impose some additional conditions on the coefficients a_k. Here we shall require that they should be periodic with period T_0 as functions of t for every fixed value of τ. This condition can be justified by the following argument. For $\alpha = 0$ the solutions are obviously periodic with period T_0. Therefore, it is natural to expect that for small nonzero values of α they resemble T_0-periodic functions with slowly varying amplitudes and initial phases, this slow variations being expressed by means of dependence of the solution on τ. The ultimate justification of these additional requirements is revealed in the course of calculations since it is impossible to satisfy the given equation when the conditions imposed in the process of solution are incorrect.

From formula (73) we receive

$$\dot{x} = \sum_{k=0}^{\infty} [(a_k)'_t \alpha^k + (a_k)'_\tau \alpha^{k+1}]$$

$$\ddot{x} = \sum_{k=0}^{\infty} [(a_k)''_{tt} \alpha^k + 2(a_k)''_{t\tau} \alpha^{k+1} + (a_k)''_{\tau\tau} \alpha^{k+2}]$$

Let us substitute these expressions into (34) and impose one more condition that the equation thus obtained is satisfied identically with respect to α for any fixed values of t and τ. This is of course an additional requirement which is not implied by the conditions of the problem; to understand this, note that for the correct equality $\tau + 0 \cdot \alpha = 0 + t \cdot \alpha$ this requirement is not fulfilled. But if we manage to determine all the quantities in such a way that this condition is fulfilled equation (34) will be satisfied automatically. Equating the coefficients in the same powers of α we obtain

$$(a_0)''_{tt} + \omega_0^2 a_0 = 0$$

$$(a_1)''_{tt} + \omega_0^2 a_1 = -2(a_0)''_{t\tau} + f_1(a_0, (a_0)'_t, t) \tag{74}$$

The first relation yields

$$a_0(t, \tau) = A_0(\tau) \cos \omega_0 t + B_0(\tau) \sin \omega_0 t \tag{75}$$

Substituting this expression into the right-hand side of (74) we obtain a function which is periodic in t with period T_0 for every fixed value of τ. Since we need a periodic solution a_1 with period T_0 as function of t, the corresponding conditions eliminating resonance must hold. After some simple transformations (let the reader carry

out the calculations) these conditions take the form

$$\left.\begin{aligned} A_0'(\tau) &= -\frac{1}{2\pi}\int_0^{T_0} f_1(A_0(\tau)\cos\omega_0 t + B_0(\tau)\sin\omega_0 t, \\ &\quad -\omega_0 A_0(\tau)\sin\omega_0 t + \omega_0 B_0(\tau)\cos\omega_0 t,\ t)\sin\omega_0 t\,dt \\ B_0'(\tau) &= \frac{1}{2\pi}\int_0^{T_0} f_1(A_0(\tau)\cos\omega_0 t + B_0(\tau)\sin\omega_0 t, \\ &\quad -\omega_0 A_0(\tau)\sin\omega_0 t + \omega_0 B_0(\tau)\cos\omega_0 t,\ t)\cos\omega_0 t\,dt \end{aligned}\right\} \tag{76}$$

We have thus obtained a system of two ordinary differential equations with respect to the two unknown functions $A_0(\tau)$ and $B_0(\tau)$ from which these functions can be determined. After $A_0(\tau)$ and $B_0(\tau)$ have been found we can write, by virtue of (74) (since there is no resonance), the expression

$$a_1(t,\ \tau) = \varphi_1(t,\ \tau) + A_1(\tau)\cos\omega_0 t + B_1(\tau)\sin\omega_0 t$$

where φ_1 is a completely determined function with period T_0 in t for every fixed τ and the functions $A_1(\tau)$ and $B_1(\tau)$ are yet unknown. Taking the equation for a_2 (following equation (74)) and again making use of the conditions excluding resonant terms we obtain a new system of two ordinary differential equations in $A_1(\tau)$ and $B_1(\tau)$, etc. The process can be, in principle, continued indefinitely.

The system of equations (76) can also be obtained on the basis of the following argument. Let us write equation (34) in the form of the system

$$\dot{x} = \omega_0 y,\quad \dot{y} = -\omega_0 x + \frac{\alpha}{\omega_0} f_1(x,\ \omega_0 y,\ t) + \ldots \tag{77}$$

For $\alpha = 0$ the corresponding flow in the phase plane x, y coincides with the uniform rotation of the plane in negative direction with angular speed ω_0. Therefore, it is natural to consider the solution of equation (77) for small $\alpha \neq 0$ in a uniformly rotating coordinate system. This means that we make the change of the unknown functions according to the formulas

$$x = \tilde{x}\cos\omega_0 t + \tilde{y}\sin\omega_0 t,\quad y = -\tilde{x}\sin\omega_0 t + \tilde{y}\cos\omega_0 t \tag{78}$$

where $\tilde{x}$ and $\tilde{y}$ are not constant, as in the case $\alpha = 0$, but are functions of t, their rate of change naturally being small for small values of $|\alpha|$. The substitution of (78) into (77) results in the equation (verify the computations!)

$$\frac{d\tilde{x}}{dt} = -\left[\frac{\alpha}{\omega_0} f_1(\tilde{x}\cos\omega_0 t + \tilde{y}\sin\omega_0 t, -\omega_0\tilde{x}\sin\omega_0 t + \omega_0\tilde{y}\cos\omega_0 t,\ t) + \ldots\right]\sin\omega_0 t \tag{79}$$

and in an analogous equation for $\tilde{y}$. When $|\alpha|$ is small the right member of (79) is close to a periodic function since $\tilde{x}$ and $\tilde{y}$ are "almost constant". For instance, the rate of change of $\tilde{x}$ is proportional to α and its systematic component is equal to a small constant term in Fourier's expansion of the right member of (79) while all the other Fourier components result in a vibration with small amplitude and period T_0 about this principal component. To obtain this principal term we must average the right-hand side of (79) (e.g. see [107], Sec. XVII.23) and discard the terms of higher order of smallness, which directly leads (check it up!) to the first equations (76). This accounts for the term the **method of averaging** applied to this procedure. In this form the method can also be applied to equations not involving a small parameter, but this procedure should be combined with an estimation of the errors in the computational process (in the case of forced vibrations the quantity ω_0 in (78) should be understood as the frequency of the external action).

Let the reader check that the same result is obtained if we substitute into (34) the expression $x = \tilde{x}\cos\omega_0 t + \tilde{y}\sin\omega_0 t$, discard the terms with $\frac{d^2\tilde{x}}{dt^2}$ and $\frac{d^2\tilde{y}}{dt^2}$, expand all the terms into Fourier's series, and then equate the coefficients in $\cos\omega_0 t$ and in $\sin\omega_0 t$.

If we are only interested in periodic solutions $x(t)$ with period T_0 the coefficients a_j must be independent of τ, and therefore the coefficients A_j and B_j should be constant. Under this condition equations (76) become determining equations (37), and thus we return to the method described in Sec. 4. In the general case we deal with nonconstant solutions of system (76) yielding nonperiodic solutions of the original system. Note that (76) is an autonomous system and therefore can be investigated on the $A_0\ B_0$-plane with the aid of the methods described in Secs. 2.3-6. The rest points in the phase plane determine periodic solutions of equation (34), the generated vibrations having the frequency of the external excitation. In such a case we speak of the so-called **synchronization**. To cycles in the phase plane A_0, B_0 there correspond quasi-periodic solutions whose frequencies result from the superposition of the frequency ω_0 and the frequency of the cycle (these are the so-called **combination frequencies**). The first frequency being considerably greater than the second, there arise beats in such a vibration. A trajectory in the A_0B_0-plane entering a rest point specifies transient processes passing into a stable periodic oscillation.

The simplest case occurs when the function f_1 is independent of t. Then we can introduce, by means of the formulas $A_0 = \rho\cos\varphi$, $B_0 = \rho\sin\varphi$, the polar coordinates ρ, φ in the A_0B_0-plane, which

yields

$$\frac{d\rho}{dt}=\frac{1}{\rho}(A_0A_0'+B_0B_0')=$$

$$=-\frac{1}{2\pi}\int_0^{T_0} f_1(\rho\cos(\omega_0 t-\varphi),\ -\omega_0\rho\sin(\omega_0 t-\varphi))\sin(\omega_0 t-\varphi)\,dt=$$

$$=-\frac{1}{2\pi\omega_0}\int_0^{2\pi} f_1(\rho\cos\psi,\ -\omega_0\rho\sin\psi)\sin\psi\,d\psi$$

Thus we have obtained a differential equation of the first order with one unknown function solvable by separation of variables. After $\rho(\tau)$ has been found the function $\varphi(\tau)$ is determined from the equation

$$\frac{d\varphi}{d\tau}=\frac{1}{\rho^2}(A_0B_0'-B_0A_0')=\frac{1}{2\pi\omega_0\rho}\int_0^{2\pi} f_1(\rho\cos\psi,\ -\omega_0\rho\sin\psi)\cos\psi\,d\psi$$

by direct integration.

The method described above can be applied to more general equations than (34) involving functions f_j dependent on τ. In this case a system of type (76) obtained is nonautonomous but nevertheless in many problems it is simpler than the original equation. By the way, the system for determining A_1 and B_1 involves, in the general case, the functions $A_0(\tau)$ and $B_0(\tau)$, and is nonautonomous even for system (34) unless we consider the equations for A_0, B_0, A_1, and B_1 as a system of simultaneous equations in the four-dimensional phase space of the variables A_0, B_0, A_1, B_1.

If we deal with a stable vibration, system (76) can be used for determining approximate values of $A_0(\infty)$ and $B_0(\infty)$ by means of numerical integration, these quantities specifying the amplitude and the phase of the vibration in the zeroth approximation. The numerical integration can be combined with Newton's method: the former is used for finding some initial approximations after which the latter is applied to the system of determining equations in order to obtain a more accurate result.

Let us apply this technique to forced oscillations, in a system with self-excitation of type (56), described by the equation

$$\ddot{x}+\omega_0^2x=\alpha(\dot{x}-k\dot{x}^3)+\alpha M\cos\omega t \tag{80}$$

where $\omega=\omega_0+\omega_1\alpha$. As in Sec. 4, we make the change $\omega_0=\omega-\omega_1\alpha$, which yields

$$\ddot{x}+\omega^2x=(2\omega\omega_1x+\dot{x}-k\dot{x}^3+M\cos\omega t)\,\alpha-\omega_1^2x\alpha^2 \tag{81}$$

According to (73), we take the expansion

$$x=a_0(t,\ \tau)+a_1(t,\ \tau)\,\alpha+\ldots \qquad (\tau=\alpha t)$$

and then, as in the general case, obtain the expression

$$a_0 = A_0(\tau)\cos\omega t + B_0(\tau)\sin\omega t \tag{82}$$

Equating in (81) the coefficients in α we receive

$$\frac{\partial^2 a_1}{\partial t^2} + \omega^2 a_1 = 2\omega_1\omega a_0 - 2\frac{\partial^2 a_0}{\partial t\,\partial\tau} + \frac{\partial a_0}{\partial t} - k\left(\frac{\partial a_0}{\partial t}\right)^3 + M\cos\omega_0 t =$$

$$= 2\omega_1\omega(A_0\cos\omega t + B_0\sin\omega t) + 2\omega(A_0'\sin\omega t - B_0'\cos\omega t) -$$

$$-\omega(A_0\sin\omega t - B_0\cos\omega t) + k\omega^3(A_0\sin\omega t - B_0\cos\omega t)^3 + M\cos\omega t$$

The conditions of the absence of resonance give us the system of equations

$$\left.\begin{aligned}\frac{dA_0}{d\tau} &= \frac{1}{2}A_0 - \omega_1 B_0 - \frac{3}{8}k\omega^2(A_0^3 + A_0B_0^2)\\ \frac{dB_0}{d\tau} &= \frac{1}{2}B_0 + \omega_1 A_0 - \frac{3}{8}k\omega^2(B_0^3 + B_0A_0^2) + \frac{M}{2\omega}\end{aligned}\right\} \tag{83}$$

(let the reader verify the result). Equality (82) can be rewritten as $a_0 = A_0(\alpha t)\cos\omega t + B_0(\alpha t)\sin\omega t$ and equations (83) are the autonomous system whose trajectories specify the sought-for forced oscillations in the zeroth approximation. A more thorough analysis (for which we refer the reader to [124]) shows that for small values of ω_1 the trajectories of system (83) enter rest points, that is, in the limit (as $t \to \infty$), the solutions of equation (80) are periodic; if ω_1 assumes large values system (83) possesses a stable limit cycle, that is there arise vibrations with combination frequencies.

As in Sec. 5, in the case of functions f_j dependent on t periodically with period $\frac{T_0}{m}$ $(m = 2, 3, \ldots)$ the above construction leads to subharmonic oscillations (in this connection see [55]).

The averaging method can also be applied to systems with several degrees of freedom. For instance, generalizing equation (79), we can take a system of the form

$$\dot{\mathbf{x}} = \alpha\mathbf{f}(x, t, \alpha) \tag{84}$$

whose right-hand side is periodic or almost periodic (Sec. 1.7) in t. Denote by $\bar{\mathbf{f}}(x)$ the average value of the function $\mathbf{f}|_{\alpha=0}$ taken with respect to t:

$$\bar{\mathbf{f}}(x) = \lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbf{f}(x, t, 0)\,dt \tag{85}$$

Let us make the change of the unknown functions according to the formula

$$\mathbf{x} = \mathbf{y} + \alpha\left[\int_0^t \mathbf{f}(y, t_1, \alpha)\,dt_1 - t\bar{\mathbf{f}}(y)\right] \tag{86}$$

where x and y under the sign of integration are regarded as parameters (the same of course refers to formula (85)). Taking into account that $\dot{y} = O(\alpha)$ we obtain (check it up!) the transformed system

$$\dot{\mathbf{y}} = -\alpha \left[\int_0^t \mathbf{f}(y, t_1, \alpha)\, dt_1 - t\bar{\mathbf{f}}(y) \right] + \alpha \mathbf{f}(y, t, \alpha) + \\ + O(\alpha^2) = \alpha \bar{\mathbf{f}}(y) + O(\alpha^2)$$

Finally, discarding the terms of higher order of smallness we arrive at the averaged system corresponding to system (84):

$$\dot{\mathbf{y}} = \alpha \bar{\mathbf{f}}(y) \tag{87}$$

System (87) is autonomous and, of course, simpler than (84). Many problems related to this system, such as the determination of the type of its rest points and the investigation of their stability, can be readily solved in simpler cases. It can be shown that to each rest point of system (87) there corresponds a periodic (almost periodic) solution of system (84) with periodic (almost periodic) right-hand side which is located in an α-neighbourhood of the rest point. If the rest point is stable the corresponding solution of system (84) is stable as well. On the construction of higher approximations by the method of averaging see [91].

The scheme of the averaging method can be applied to the system

$$\dot{\mathbf{x}} = \alpha \mathbf{f}(x, \theta, t, \alpha), \quad \dot{\theta} = (\lambda_1, \lambda_2, \ldots, \lambda_p)^* + \alpha \varphi_1(x, \theta, t, \alpha) \tag{88}$$

which is a generalization of system (84). Here $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ is a collection of angular variables (phases) on which all the functions f_j and φ_j depend periodically. The corresponding averaged system has the form

$$\left.\begin{aligned} \dot{\mathbf{y}} &= \alpha \lim_{T \to \infty} \frac{1}{T} \int_0^T \mathbf{f}(y, t\lambda, t, 0)\, dt \\ \dot{\chi} &= \lambda + \alpha \lim_{T \to \infty} \frac{1}{T} \int_0^T \varphi_1(y, t\lambda, t, 0)\, dt \end{aligned}\right\}$$

An important feature of this averaged system is that the variables y are found separately. In this case the solutions of the averaged system also approximate the solutions of the original system to within the terms of higher order of smallness.

In particular, the so-called **system with rapidly rotating phases** written in the form

$$\dot{\mathbf{x}} = \mathbf{f}(x, \theta, t), \quad \dot{\theta} = M(\lambda_1, \lambda_2, \ldots, \lambda_p)^* + \varphi(x, \theta, t) \tag{89}$$

is reducible to (88). Here M is a large parameter, and the system is considered for $M \to \infty$. After the "**fast time**" $s = Mt$ and the small parameter $\alpha = M^{-1}$ have been introduced, system (89) turns into the system

$$\frac{dx}{ds} = \alpha \mathbf{f}(x, \theta, \alpha s), \quad \frac{d\theta}{ds} = (\lambda_1, \lambda_2, \ldots, \lambda_p)^* + \alpha \varphi(x, \theta, \alpha s)$$

belonging to type (88).

11. Krylov-Bogolyubov Method. In 1937 N.M. Krylov and N.N. Bogolyubov inaugurated one of the most general and flexible asymptotic schemes for expanding solutions in series with respect to powers of a small parameter entering into the equations. We shall demonstrate this method by the example of the autonomous equation (47). In this case the solution is looked for in the form

$$x = a \cos \psi + x_1(a, \psi)\alpha + x_2(a, \psi)\alpha^2 + \ldots (a = a(t, \alpha), \psi = \psi(t, \alpha)) \quad (90)$$

where all the functions $x_j(a, \psi)$ to be determined are periodic with period 2π with respect to ψ, the functions a and ψ being supposed to satisfy equations of the form

$$\dot{a} = A_1(a)\alpha + A_2(a)\alpha^2 + \ldots, \quad \dot{\psi} = \omega_0 + B_1(a)\alpha + B_2(a)\alpha^2 + \ldots \quad (91)$$

whose coefficients are found in the process of the construction of the solution. (Let the reader verify that the chosen form of the solution is coherent with the original system for $\alpha = 0$.) The functions x_j, A_j and B_j should be constructed in such a way that after the derivatives $\dot{x}(a, \psi)$ and $\ddot{x}(a, \psi)$ have been found from (90) and (91) by means of formal differentiation and substituted into equation (47), the latter must be satisfied identically with respect to α for any fixed a and ψ (this is of course sufficient for the original equation to be satisfied but not necessary since a and ψ are themselves functions of α). Besides, to normalize the expressions obtained we require that Fourier's expansions of the functions x_j should not contain terms with $\cos\psi$ and $\sin\psi$ (the justification of this requirement lies in the fact that such terms can be combined with the first term on the right-hand side of (90)).

Let us demonstrate the technique for determining the first approximation. Computing $\dot{x}$ and $\ddot{x}$ and discarding the terms with α^2 we obtain

$$\dot{x} = \dot{a}\cos\psi - a\sin\psi \cdot \dot{\psi} + \left(\frac{\partial x_1}{\partial a}\dot{a} + \frac{\partial x_1}{\partial \psi}\dot{\psi}\right)\alpha + \ldots =$$

$$= A_1(a)\cos\psi \cdot \alpha - a\sin\psi \cdot \omega_0 - B_1(a)\, a \sin\psi \cdot \alpha \frac{\partial x_1}{\partial \psi}\omega_0\alpha + \ldots,$$

$$\ddot{x} = -2\omega_0 A_1(a)\sin\psi \cdot \alpha - a\cos\psi \cdot \omega_0^2 - 2B_1(a)\, a\omega_0 \cos\psi \cdot \alpha +$$

$$+ \frac{\partial^2 x_1}{\partial \psi^2}\omega_0^2\alpha + \ldots,$$

Substituting these expressions into (47) and equating the terms involving α we arrive at the equality

$$\omega_0^2\left(\frac{\partial^2 x_1}{\partial\psi^2}+x_1\right)=f_1(a\cos\psi_1,-a\omega_0\sin\psi)+2\omega_0 A_1(a)\sin\psi+$$
$$+2a\omega_0 B_1(a)\cos\psi \qquad (92)$$

The function x_1 must be periodic in ψ, and therefore the right member of (92) is subject to conditions excluding resonant terms. These conditions reduce to the equalities

$$A_1(a)=-\frac{1}{2\pi\omega_0}\int_{-\pi}^{\pi} f_1(a\cos\psi,\ -a\omega_0\sin\psi)\sin\psi\,d\psi$$

and

$$B_1(a)=-\frac{1}{2\pi a\omega_0}\int_{-\pi}^{\pi} f_1(a\cos\psi,\ -a\omega_0\sin\psi)\cos\psi\,d\psi$$

On condition that these equalities are fulfilled the right-hand side of (92) has Fourier's expansion of the form

$$g_0(a)+\sum_{j=2}^{\infty}[g_j(a)\cos j\psi+h_j(a)\sin j\psi]$$

whence, according to the normalization condition stated above, we find

$$x_1(a,\ \psi)=\frac{g_0(a)}{\omega_0^2}-\frac{1}{\omega_0^2}\sum_{j=2}^{\infty}\frac{1}{j^2-1}[g_j(a)\cos j\psi+h_j(a)\sin j\psi]$$

Thus, the expression

$$x=a\cos\psi+x_1(a,\ \psi)\alpha \qquad (93)$$

we have constructed in which a and ψ are determined by the equations

$$\dot a=A_1(a)\alpha,\quad \dot\psi=\omega_0+B_1(a)\alpha \qquad (94)$$

satisfies equation (47) with discrepancy of the order of α^2. We can therefore conclude that if the approximation we have found satisfies the same initial conditions as the exact solution (this can be achieved by choosing appropriate values of the arbitrary constants appearing in the integration of equations (94), which is readily carried out), the difference between the approximate solution and the exact one is of the order of $\alpha^2 t$. The construction of higher approximations is performed in a similar way but involves more complicated calculations.

If the approximate solution is considered on a time interval whose length is of the order of $\frac{1}{\alpha}$ (and thus tends to infinity as $\alpha \to 0$) the error of the approximation is of the order of α (why?), and therefore it makes no sense to write the second summand in formula (93). Consequently, the approximate solution should be taken in the simple form $a\cos\psi$ where a and ψ are determined from equation (94). This yields exactly the same result as if we applied the method of Sec. 10 to the original equation after the independent variable is changed as in Sec. 7 (check it up!).

If the parameters of the original system slowly vary, that is if the right-hand side of equation (47) and the coefficient ω_0^2 depend on the slow time $\tau = \alpha t$ (Sec. 10), all the functions x_j, A_j and B_j we construct must also depend on τ. (Let the reader analyse in what way the formulas for the first approximation are changed in this case and show that for the equation $\frac{d}{dt}[m(\tau)\dot{x}] + k(\tau)x = f(x, \tau)\alpha$ the function a in the first approximation $x = a\cos\psi$ is expressed by the formula $a = \text{const}\,[m(\tau)k(\tau)]^{-\frac{1}{4}}$; also show that for the equation $\frac{d}{dt}[m(\tau)\dot{x}] + k(\tau)x = f(\dot{x}, \tau)\alpha$ we obtain $\psi = \int \sqrt{\frac{k(\tau)}{m(\tau)}}\,dt$, $\tau = \alpha t$.)

Now let us consider the equation of forced oscillations of form (34) in which all the functions f_j are periodic in t with period T, the periods T and $T_0 = \frac{2\pi}{\omega_0}$ being incommensurable (this means that we deal with a nonresonant case). The solution of this equation can be constructed in the form

$$x = a\cos\psi + x_1(a, \psi, t)\alpha + x_2(a, \psi, t)\alpha^2 + \ldots$$

where all the functions x_j are 2π-periodic in ψ and T-periodic in t, the amplitude a and the phase ψ satisfying equations (91). The construction of the functions x_j is performed in just the same way as in the autonomous case but involves expansions in Fourier's double series (e.g. see [107], Sec. XVII.30) with respect to ψ and t. Let the reader derive the formulas of the first approximation for this case.

If we have resonance, that is $T = \frac{p}{q}T_0$ where p and q are integers, it is more convenient to construct the solution in the form

$$x = a\cos(\omega_0 t + \theta) + x_1(a, \theta, t)\alpha + x_2(a, \theta, t)\alpha^2 + \ldots$$

where all the functions x_j are periodic in t with period qT and the quantities a and θ satisfy the equations

$$\dot{a} = A_1(a, \theta)\alpha + A_2(a, \theta)\alpha^2 + \ldots, \quad \dot{\theta} = B_1(a, \theta)\alpha + B_2(a, \theta)\alpha^2 + \ldots$$

whose right-hand sides, as in the case of equations (91), are determined in the process of constructing the solution. For more detail on this rather complicated construction we refer the reader to [12]. We only remark that the series in powers of the small parameter α appearing here are, generally speaking, only asymptotically convergent (Sec. II.4.1). Practically, however, this is not essential since we usually take only approximations of low orders, most often of the first order. For practical purposes it is more important to find an upper bound for the values of the parameter α for which the obtained asymptotic formulas can be used. This question is usually answered on the basis of the results of the numerical integration of some simpler equations.

In [12] the reader can also find a description of the application of the above method to investigating monofrequent oscillations in systems with several degrees of freedom. In this case, instead of (90), the expansion of the solution is taken in the form

$$\mathbf{x} = (a\cos\psi)\,\boldsymbol{\varphi} + \mathbf{x}_1(a,\ \psi)\,\alpha + \mathbf{x}_2(a,\ \psi)\,\alpha^2 + \ldots$$

where $\boldsymbol{\varphi}$ is the corresponding eigenvector for $\alpha = 0$ and the quantities a and ψ satisfy scalar equations (91). This construction only yields a two-parameter family of particular solutions while the general solution may involve more than two arbitrary constants. But nevertheless this result is important in some cases when the two-parameter family found is stable and attracts all the other solutions passing sufficiently close to it.

12. Discrete-Time Systems. In problems of the theory of nonlinear vibrations and in some other problems the so-called discrete-time systems play an important role. A system of this type is defined by a two-way sequence of homeomorphisms (that is of continuous in both directions one-to-one mappings)

$$x' = \psi_k(x) \qquad (k = \ldots,\ -2,\ -1,\ 0,\ 1,\ 2,\ \ldots) \tag{95}$$

of the n-dimensional space of n-tuples $x = (x_1, x_2, \ldots, x_n)$ into itself. The values of the number k play the role of time moments, and formula (95) determines a law according to which the system instantaneously passes, "jumps", from one state to another at these time moments while for the intermediate time moments it is regarded as being in a state of rest.

To determine the law of motion of a separate point we must set its position for the time interval $k_0 - 1 \leqslant t < k_0$ before it jumps at the moment $t = k_0$. After this jump the point is in the position $\psi_{k_0}(x_0)$ until the next jump at the moment $t = k_0 + 1$ is made; on the time interval $k_0 + 1 \leqslant t < k_0 + 2$ the point occupies the position $\psi_{k_0+1}(\psi_{k_0}(x_0))$ until it jumps at the moment $t = k_0 + 2$ and so on. We can also use inverse mappings to consider time moments preceding $t = k_0$; then the position x_0 appears after the jump, made

at the time moment $t = k_0 - 1$, from the position $\psi_{k_0-1}^{-}(x_0)$ where the superscript "—" designates the inverse mapping, etc. Hence, we obtain a two-way sequence

$$x = x(k, k_0, x_0) \qquad (k = \ldots, -2, -1, 0, 1, 2, \ldots) \tag{96}$$

specifying the law of motion of the point. The function $x(k, k_0, x_0)$ possesses the following obvious properties:

$$\psi_k(x(k, k_0, x_0)) = x(k+1, k_0, x_0), \qquad x(k_0, k_0, x_0) = x_0 \tag{97}$$

When mappings (95) are independent of k we obtain a simpler case in which we say, by the analogy with autonomous first-order systems of differential equations, that the discrete-time system in question is autonomous. Then the position of the point moving in a jump-like manner depends solely on its initial position and on the number of jumps. This means that instead of (96) and (97) we have the formulas

$$x = x(k-k_0, x_0), \quad \psi(x(k, x_0)) = x(k+1, x_0), \quad x(0, x_0) = x_0$$

By analogy with Sec. 2.1, we can interpret this motion as a discrete flow in the space E_n for which the role of trajectories is played by two-way sequences of points. As in Sec. 2.1, we can classify the "trajectories" into the following three types: nonclosed trajectories (that is two-way sequences consisting of an infinite number of distinct points), closed ones (consisting of a finite number of different points repeating in a cyclic order) and "rest points" determined by the equation $\psi(x) = x$. (To achieve a closer analogy with differential equations we can write the mapping $x' = \psi(x)$ in the form $\mathbf{x}' = \mathbf{x} + \mathbf{f}(x)$; then the rest points are found from the equation $\mathbf{f}(x) = \mathbf{0}$.) Following Sec. 2.2, we can speak about limit sets of trajectories; all the properties enumerated in Sec. 2.2 are readily extended to this case except, of course, Property 5.

In the investigation of the behaviour of a trajectory in the vicinity of a rest point (for definiteness, let it be at the origin) a fundamental role is played by the trajectories of the linearized system

$$\mathbf{x}' = \mathbf{A}\mathbf{x} \tag{98}$$

where $\mathbf{A}$ is a square matrix of order n. The law of motion for such a system is quite simple:

$$\mathbf{x}(k, \mathbf{x}_0) = \mathbf{A}^k \mathbf{x}_0$$

It follows immediately that for the zero solution of system (98) (and therefore for any solution as well) to be asymptotically stable it is necessary and sufficient that the moduli of all the eigenvalues of the matrix $\mathbf{A}$ be less than unity. If at least one of the moduli exceeds unity the zero solution is unstable. For a more thorough

investigation it is advisable to introduce in E_n a new affine coordinate system connected with the old system by an equality $\mathbf{x}=\mathbf{Ty}$. In the new coordinates the equation of the mapping (98) is transformed to $\mathbf{y}'=\mathbf{T}^{-1}\mathbf{ATy}$, and the matrix $\mathbf{T}$ can be chosen in such a way that the new matrix $\mathbf{T}^{-1}\mathbf{AT}$ specifying the mapping takes a simpler form (for instance, the Jordan canonical form; see Sec. IV.3.4). We suggest that, by analogy with Sec. 2.3, the reader give a complete classification of the isolated rest points of a linear autonomous system with discrete time in the plane. As in Sec. 2.4, it turns out that the type of a rest point remains the same when we return from the linearized system to the original one except a point of the type of the centre for which the matrix $\mathbf{A}$ possesses eigenvalues whose moduli are equal to unity. In the case of a centre mapping (98) reduces, after the corresponding affine transformation of the plane has been made, to a rotation of the plane about the origin through an angle πs (where $e^{\pm i\pi s}$ is an eigenvalue of the matrix $\mathbf{A}$). Therefore, for rational values of s all the trajectories except the rest point itself are cycles whereas for irrational values of s the trajectories cover a set which is everywhere dense in an elliptic region.

Let us proceed to study nonhomogeneous linear discrete systems with constant coefficients:

$$\mathbf{x}'=\mathbf{Ax}+\mathbf{a}_k \tag{99}$$

It can be easily verified that if the initial condition is set for $k=0$ the solution has, for $k>0$, the form

$$\mathbf{x}_1=\mathbf{Ax}_0+\mathbf{a}_0,\ \mathbf{x}_2=\mathbf{A}^2\mathbf{x}_0+\mathbf{Aa}_0+\mathbf{a}_1,\ \ldots$$

and, generally,

$$\mathbf{x}_k=\mathbf{x}(k,\ \mathbf{x}_0)=\mathbf{A}^k\mathbf{x}_0+\mathbf{A}^{k-1}\mathbf{a}_0+\mathbf{A}^{k-2}\mathbf{a}_1+\ldots+\mathbf{a}_{k-1}$$

(Write down the solution for $k<0$.)

Now let us consider the particular case when the sequence $\mathbf{a}_k$ is periodic with period m, i.e. $\mathbf{a}_{k+m}\equiv\mathbf{a}_k$. Then, for an initial point $\mathbf{x}_0$ to determine a periodic solution with the same period, it is necessary and sufficient that $\mathbf{x}_m=\mathbf{x}_0$, which means that

$$(\mathbf{A}^m-\mathbf{I})\,\mathbf{x}_0=-\mathbf{A}^{m-1}\mathbf{a}_0-\mathbf{A}^{m-2}\mathbf{a}_1-\ldots-\mathbf{a}_{m-1}$$

This equation possesses a unique solution if and only if

$$\det(\mathbf{A}^m-\mathbf{I})\neq 0 \tag{100}$$

i.e. if none of the values of the root $\sqrt[m]{1}$ is an eigenvalue of the matrix $\mathbf{A}$. If this condition is fulfilled it is convenient to use the representation

$$\mathbf{a}_k=\sum_{p=0}^{m-1}\exp\left(\frac{2\pi i}{m}pk\right)\mathbf{c}_p\ (k=0,\ \pm1,\pm2,\ \ldots) \tag{101}$$

for finding the periodic solution. Representation (101) is analogous to Fourier's series and is possible for every periodic sequence $\mathbf{a}_k$ with period m. The coefficients of this representation are found by the formula

$$\mathbf{c}_p = \frac{1}{m} \sum_{k=0}^{m-1} \exp\left(-\frac{2\pi i}{m} pk\right) \mathbf{a}_k \quad (p = 0, 1, \ldots, m-1) \qquad (102)$$

(Check up this formula by substituting (102) into (101). Prove the uniqueness of representation (101).) If we look for an m-periodic solution $\mathbf{x}$ of system (99) as an expansion of form (101) with undetermined coefficients $\mathbf{d}_p$ (replacing $\mathbf{c}_p$) simple computations which we leave to the reader lead to the formula

$$\mathbf{x} = -\sum_{p=0}^{m-1} \exp\left(\frac{2\pi i}{m} pk\right) \left[\mathbf{A} - \left(\exp\frac{2\pi i}{m} p\right) \mathbf{I}\right]^{-1} \mathbf{c}_p$$

Thus, *if condition* (100) *holds system* (99) *possesses an m-periodic solution for every periodic sequence* $\mathbf{a}_k$ *with period m.* It can also be proved that *this solution is unique.* But if for some $p = p_0$ the number $\exp\left(\frac{2\pi i}{m} p_0\right)$ is an eigenvalue of the matrix $\mathbf{A}$ then, generally speaking, for $\mathbf{c}_{p_0} \neq \mathbf{0}$ resonance sets in the system, that is there is no periodic solution.

Systems (95) and (99) are similar to systems of differential equations of the first order. One can easily write down discrete analogues of differential equations and systems of differential equations of higher order. For instance, system (99) can be written in the form $\mathbf{x}_{k+1} - \mathbf{A}\mathbf{x}_k + \mathbf{a}_k$, and hence the analogue of a scalar differential equation of the second order should be written in the form

$$x_{k+2} + ax_{k+1} + bx_k = a_k \qquad (103)$$

and so on. Thus, we arrive at difference equations mentioned in Sec. III.2.3 (also see [107], Sec. XVII.16). From a difference equation and from a system of any order it is possible to pass to a system of the first order. For example, if we put $x_{k+1} = y_k$ in equation (103) we obtain the system

$$x_{k+1} = y_k, \quad y_{k+1} = -bx_k - ay_k + a_k \qquad (104)$$

of two equations of the first order. In particular, it follows that the above assertions concerning periodic solutions can readily be extended to equation (103), which, by the way, can be proved directly without passing to system (104). In this case, instead of condition (100), we must introduce the requirement that none of the values of the root $\sqrt[m]{1}$ should be a solution of the characteristic equation

$$\lambda^2 + a\lambda + b = 0$$

Consider the equation

$$x_{k+2}-2\left(\cos\frac{2\pi p_0}{m}\right)x_{k+1}+x_k=a_k \qquad (105)$$

with real a_k which is a particular case of equation (103). If $\cos\frac{2\pi p_0}{m}=\pm 1$, one of the characteristic roots of equation (105) is equal to ± 1. If $\cos\frac{2\pi p_0}{m}\neq\pm 1$ the characteristic roots are $\lambda_{1,2}=$ $=\exp\left(\pm\frac{2\pi i}{m}p_0\right)$. In this case equation (105) is an analogue of the differential equation $\ddot{x}+\omega_0^2 x=f(t)$.

This far-reaching analogy between systems with discrete and continuous time makes it possible, in particular, to extend to the former most of the methods described in this chapter. Let the reader analyse such an extension.

In conclusion we mention that discrete-time systems can be applied to investigating periodic systems (generally speaking, nonlinear ones) with continuous time, for instance, the differential equation

$$\dot{\mathbf{x}}=\mathbf{f}(x,\ t) \qquad (106)$$

with a periodic right-hand side of period T_0 in the variable t. This is performed in the following way. We arbitrarily fix a value $t=t_0$ and then associate with every point $x_0\in E_n$ the value (denote it as $\psi(x_0)$) assumed for $t=t_0+T_0$ by the solution $x(t,\ t_0,\ x_0)$ of system (106) satisfying the initial condition $x|_{t=t_0}=x_0$. System (106) being periodic, the passages from t_0+T_0 to t_0+2T_0, from t_0+2T_0 to t_0+3T_0, etc. determine a mapping ψ of the type considered above; it is sometimes called a shift operator along the trajectories of system (106) for one period. (It should be noted that this mapping depends on the choice of t_0 but one can easily show that this does not affect the possibility of its application indicated below.) Thus, putting $x_k=x(t_0+kT_0,\ t_0,\ x_0)$ we arrive at the autonomous discrete-time system $x_{k+1}=\psi(x_k)$ to which the methods of investigation of autonomous systems can be applied and, in particular, the method of the phase space. This approach to the analysis of T_0-periodic system (106) is known as the **stroboscopic method.** When it is applied we only consider the values assumed by the solutions of system (106) at the discrete time moments t_0+kT_0. In this construction, to a rest point of the resultant discrete-time system there corresponds a periodic solution of system (106) with period T_0, to a cycle of the discrete system there corresponds a subharmonic oscillation of system (106), and so on. The practical application of this method encounters a difficulty connected with the passage from (106) to the corresponding discrete system since this involves the general solution of (106). This difficulty can sometimes be overcome by means of the small-parameter method, numerical integration, etc. The method also proves useful for theoretical aspects of periodic solutions of differential equations.

Bibliography *

1. Aizerman, M. A., Gantmacher, F. R., *Absolute Stability of Regulating Systems*, Moscow, Nauka, 1965 (Russian edition).
2. Akhiezer, N. I., *Lectures on Calculus of Variations*, Moscow, Gostekhizdat, 1955 (Russian edition). The English translation: *The Calculus of Variations*, New York, Blaisdell Publishing Company, 1962.
3. Akhiezer, N. I., *Elements of the Theory of Elliptic Functions*, 2nd edition, Moscow, Nauka, 1970 (Russian edition).
4. Andronov, A. A., Vitt, A. A., Khaikin, S. E., *The Theory of Vibrations*, Moscow, Fizmatgiz, 1959 (Russian edition).
5. Angot, A., Compléments de mathématiques. A l'usage des ingéniurs de l'électrotechnique et des télécommunications, Paris, 1957.
6. Babakov, I. M., *The Theory of Vibrations*, 2nd edition, Moscow, Nauka, 1968 (Russian edition).
7. Barbashin, E. A., *An Introduction to the Theory of Stability*, Moscow, Nauka, 1967 (Russian edition).
8. Bellman, R., *Stability Theory of Differential Equations*, New York, McGraw-Hill, 1969.
9. Berezin, I. S., Zhidkov, N. P., *Computation Methods*, 3rd edition, V. 1, Moscow, Nauka, 1967 (Russian edition).
10. Blackwell, D., Girshick, M. A., *Theory of Games and Statistical Decisions*, New York, Wiley, London, Chapman and Hall, 1954.
11. Bliss, G. A., *Lectures on the Calculus of Variations*, Chicago, Univ. of Chicago, 1946.
12. Bogolyubov, N. N., Mitropolsky, Yu. A., *Asymptotic Methods in the Theory of Nonlinear Oscillations*, 3rd edition, Moscow, Fizmatgiz, 1963 (Russian edition). The English translation: *Asymptotic Methods in the Theory of Nonlinear Oscillations*, New York, Gordon Breach Science Publ., 1962.
13. Bohr, H., *Almost Periodic Functions*, New York, Chelsia Publishing Company, 1947.
14. Borisenko, A. I., Tarapov, I. E., *Vector Analysis and Fundamentals of Tensor Analysis*, 3rd edition, Moscow, Vysshaya Shkola, 1966 (Russian edition). The English translation: *Vector and Tensor Analysis with Applications*, Englewood Cliffs, N. J., Prentice-Hall, Inc., 1961.
15. Bulgakov, B. V., *Vibrations*, Moscow, Gostekhizdat, 1954 (Russian edition).
16. Carslaw, H. S., Jeager, J. C., *Operational Methods in Applied Mathematics*, London, Oxford University Press, 1948.
17. Chebotarev, N. G., *The Theory of Algebraic Functions*, Moscow, Gostekhizdat, 1948 (Russian edition).

---

* The bibliography includes those English translations of Russian books which are known to the translator.—*Tr.*

18. Chesari, L., *Asymptotic Behaviour and Stability Problems in Ordinary Differential Equations*, Berlin, Springer, 1959.
19. Chetaev, N. G., *Stability of Motion*, 3rd edition, Moscow, Nauka, 1965 (Russian edition).
20. Coddington, E. A., Levinson, N., *Theory of Ordinary Differential Equations*, New York, McGraw-Hill, 1955.
21. Copson, E. T., *Asymptotic Expansions*, Cambridge, Univ. Press, 1965.
22. Cunningham, W. J., *Introduction to Nonlinear Analysis*, New York, McGraw-Hill, 1958.
23. Davenport, W. B. Jr., Root, W. L., *An Introduction to the Theory of Random Signals and Noise*, New York-Toronto-London, McGraw-Hill, 1958.
24. De Bruijn, N. G., *Asymptotic Methods in Analysis*, Amsterdam, North-Holland Publishing Co., 1958.
25. Demidovich, B. P., Maron, I. A., *Computqtional Mathematics*, Moscow, Mir Publishers, 1973 (English edition).
26. Den Hartog, J. P., *Mechanical Vibrations*, New York, McGraw-Hill, 1956.
27. Ditkin, V. A., Kuznetsov, P. I., *Reference Book on Operational Calculus*, Moscow, Gostekhizdat, 1951 (Russian edition).
28. Ditkin, V. A., Prudnikov, A. P., *Integral Transformations and Operational Calculus*, Moscow, Fizmatgiz, 1961 (Russian edition).
29. Ditkin, V. A., Prudnikov, A. P., *Reference Book on Operational Calculus*, Moscow, Vysshaya Shkola, 1965 (Russian edition).
30. Doetsch, G., *Anleitung zum praktischen Gebrauch der Laplace-Transformation*, München, Oldenbourg, 1956.
31. Eisenhart, L. P., *Riemannian Geometry*, Princeton, Princeton Univ. Press, 1949.
32. Elsgolts, L. E., *Calculus of Variations*. Moscow, Gostekhizdat, 1952 (Russian edition).
33. Elsgolts, L. E., *Differential Equations and the Calculus of Variations*, Moscow, Mir Publishers, 1970 (English edition).
34. Erdélyi, A., *Asymptotic Expansions*, New York, Dover, 1956.
35. Erugin, N. P., *Systems of Linear Differential Equations with Periodic and Quasiperiodic Coefficients*, Minsk, AN BSSR Publishing House, 1963 (Russian edition).
36. Faddeev, D. K., Faddeeva, V. N., *Computational Methods of Linear Algebra*, 2nd edition, Moscow, Fizmatgiz, 1963 (Russian edition).
37. Filchakov, P. F., *Approximate Methods in Conformal Mappings*, Kiev, Naukova Dumka, 1964 (Russian edition).
38. Forsythe, G. E., Moler, C. B., *Computer Solution of Linear Algebraic Systems*, Englewood Cliffs, N. J., Prentice-Hall, Inc., 1967.
39. Frazer, R. A., Duncan, W. J., Collar, A. R., *Elementary Matrices and Some Applications to Dynamics and Differential Equations*, Cambridge, Univ. Press, 1950.
40. Fröman, N., Fröman, P. O. *JWKB Approximations, Contributions to the Theory*, Amsterdam, North-Holland Publishing Co., 1965.
41. Fuks, B. A., Levin, V. I., *Functions of a Complex Variable and Some of Their Applications*, Moscow, Gostekhizdat, 1951 (Russian edition). The English translation: *Functions of a Complex Variable and Some of Their Applications*, London-Paris-Frankfurt, Pergamon Press, 1961.
42. Fuks, B. A., Shabat, B. V., *Functions of a Complex Variable and Some Applications*, 3rd edition, Moscow, Nauka, 1964 (Russian edition).
43. Gakhov, F. D., *Boundary-Value Problems*, 2nd edition, Moscow, Fizmatgiz, 1963 (Russian edition).
44. Gantmacher, F. R., *The Theory of Matrices*, 2nd edition, Moscow, Nauka, 1966 (Russian edition).
45. Gantmacher, F. R., Krein, M. G., *Oscillatory Matrices and Kernels. Small Vibrations of Mechanical Systems*, Moscow, Gostekhizdat, 1950 (Russian edition).

46. Gass, S. I., *Linear Programming. Methods and Applications*, New York, McGraw-Hill, 2nd edition, 1964.
47. Gelfand, I. M., *Lectures on Linear Algebra*, 3rd edition, Moscow, Nauka, 1966 (Russian edition).
48. Gelfand, I. M., Fomin, S. V., *Calculus of Variations*, Moscow, Fizmatgiz, 1961 (Russian edition). The English translation: *Calculus of Variations*, Englewood Cliffs, N. J., Prentice-Hall, Inc., 1963.
49. Gelfand, I. M., Shilov, G. E., *Generalized Functions and Operations on Them*, Moscow, Fizmatgiz, 1958 (Russian edition). The English translation: *Generalized Functions*, V.1: Properties and Operations, New York, Academic Press, Inc., 1964.
50. Germain, P., *Mécanique des Milieux Continus*, Paris, Masson, 1962.
51. Gorelik, G. S., *Vibrations and Waves*, Moscow, Fizmatgiz, 1959 (Russian edition).
52. Hadley, G., *Nonlinear and Dynamic Programming*, Reading, 1964.
53. Hale, Jack K., *Oscillations in Nonlinear Systems*, New York, McGraw-Hill, 1963.
54. Halmos, P. R., *Lectures on Ergodic Theory*, Tokyo, The Mathematical Society of Japan, 1956.
55. Hayashi, Chihiro, *Nonlinear Oscillations in Physical Systems*, New York, McGraw-Hill, 1954.
56. Heading, J., *An Introduction to Phase-Integral Method*, London, Methuen, New York, Wiley, 1962.
57. Ince, E. L., *Ordinary Differential Equations*, New York, Dover, 1944.
58. Kantorovich, L. V., Krylov, V. I., *Approximate Methods of Higher Analysis*, 5th edition, Moscow, Fizmatgiz, 1962 (Russian edition).
59. Karacharov, K. A., Pilyutik, A. G. *An Introduction to the Theory of Technical Stability of Motion*, Moscow, Fizmatgiz, 1962 (Russian edition).
60. Karpelevich, F. I., Sadovsky, L. E., *Elements of Linear Algebra and Linear Programming*, 3rd edition, Moscow, Nauka, 1967 (Russian edition).
61. Kauderer, H., *Nichtlineare Mechanik*, Berlin, Springer, 1958.
62. Kharkevich, A. A., *Nonlinear and Parametric Phenomena in Radioengineering*, Moscow, Gostekhizdat, 1956 (Russian edition).
63. Khinchin, A. Ya., *Mathematical Foundations of Statistical Mechanics*, Moscow, Gostekhizdat, 1943 (Russian edition).
64. Kochin, N. E., *Vector Calculus and Fundamentals of Tensor Calculus*, 8th edition, Moscow, AN USSR Press, 1961 (Russian edition).
65. Kontorovich, M. I., *Operational Calculus and Nonstationary Phenomena in Electric Circuits*, Moscow, Gostekhizdat, 1953 (Russian edition).
66. Koppenfels, W., Stallmann, F., *Praxis der konformen Abbildung*, Berlin, Springer, 1959.
67. Krasnoselsky, M. A., *Positive Solutions of Operator Equations*, Moscow, Fizmatgiz, 1962 (Russian edition).
68. Krasnoselsky, M. A., *Topological Methods in the Theory of Nonlinear Integral Equations*, Moscow, Fizmatgiz, 1962 (Russian edition). The English translation: *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergamon Press, 1964.
69. Krasnoselsky, M. A., *Shift Operator along Trajectories of Differential Equations*, Moscow, Nauka, 1966 (Russian edition).
70. Krasnoselsky, M. A., Perov, A. I., Povolotsky, A. I., Zabreiko, P. P., *Vector Fields in the Plane*, Moscow, Fizmatgiz, 1963 (Russian edition).
71. Krasnoselsky, M. A., Vainikko, G. M., Zabreiko, P. P., Rutitsky, Ya. B., Stetsenko, V. Ya., *Approximate Solutions of Operator Equations*, Moscow, Nauka, 1969 (Russian edition).
72. Krasnov, M. L., Makarenko, G. I., *Operational Calculus. Stability of Motion* (Problems and Exercises), Moscow, Nauka, 1964 (Russian edition).
73. Krasovsky, N. N., *Some Problems of the Theory of Stability of Motion*,

Moscow, Fizmatgiz, 1959 (Russian edition). The English translation: *Stability of Motion*, Stanford, University Press, 1963.
74. Krylov, V. I., *Approximate Computation of Integrals*, 2nd edition, Moscow, Nauka, 1967 (Russian edition).
75. Künzi, H. P., Krelle, W., *Nichtlineare Programmierung*, Berlin, Springer, 1962.
76. Kurosh, A., *Higher Algebra*, Moscow, Mir Publishers, 1972 (English edition).
77. Lance, G. N., *Numerical Methods for High Speed Computers*, London, Iliffe, 1960.
78. La Salle, J., Lefschetz, S., *Stability by Lyapunov's Direct Method*, New York, Acad. Press, 1961
79. Lavrentyev, M. A., Lyusternik, L. A., *Fundamentals of the Calculus of Variations*, V. 1, Part 2, ONTI, Moscow, 1935 (Russian edition).
80. Lavrentyev, M. A., Lyusternik, L. A., *A Course of the Calculus of Variations*, Moscow, Gostekhizdat, 1950 (Russian edition).
81. Lavrentyev, M. A., Shabat, B. V., *Methods of the Theory of Functions of a Complex Variable*, 3rd edition, Moscow, Nauka, 1965 (Russian edition).
82. Lefschetz, S., *Differential Equations: Geometric Theory*, 2nd edition, New York, Interscience Publ., Inc., 1963.
83. Lefschetz, S., *Stability of Nonlinear Control Systems*, New York, Acad. Press, 1965.
84. Letov, A. M., *Stability of Nonlinear Regulating Systems*, 2nd edition, Moscow, Nauka, 1962 (Russian edition).
85. Levitan, B. M., *Almost Periodic Functions*, Moscow, Gostekhizdat, 1953 (Russian edition).
86. Lovitt, W. V., *Linear Integral Equations*, New York, Dover, 1950.
87. Luce, R. D., Raiffa, H., *Games and Decisions, Introduction and Critical Survey. A Study of the Behaviourial Models Project*, New York, Wiley, London, Chapman and Hall, 1957.
88. Lunts, G. L., Elsgolts, L. E., *Functions of a Complex Variable and Elements of Operational Calculus*, Moscow, Fizmatgiz, 1958 (Russian edition).
89. Lurye, A. I., *Operational Calculus with Applications to Mechanical Problems*, Moscow, Gostekhizdat, 1950 (Russian edition).
90. Lurye, A. I., *Nonlinear Problems of the Theory of Automatic Control*, Moscow, Gostekhizdat, 1951 (Russian edition).
91. Malkin, I. G., *Some Problems in the Theory of Nonlinear Oscillations*, Fizmatgiz, Moscow, 1956 (Russian edition). The English translation: *Some Problems in the Theory of Nonlinear Oscillations*, United States Atomic Energy Commission, Technical Information Service, ABC-tr-3766, 1959.
92. Malkin, I. G., *The Theory of Stability of Motion*, 2nd edition, Moscow, Nauka, 1966 (Russian edition).
93. Markushevich, A. I., *A Brief Course of the Theory of Analytic Functions*, 3rd edition, Moscow, Nauka, 1966 (Russian edition).
94. McConnell, A. J., *Application of Tensor Analysis*, New York, Dover, 1957.
95. McKinsey, J. C. C., *Introduction to the Theory of Games*, New York-Toronto-London, McGraw-Hill, 1952.
96. McLachlan, N. W., *Theory and Applications of Mathieu Functions*, Oxford, Clarendon Press, 1947.
97. Melentyev, P. V., *Approximate Calculations*, Moscow, Fizmatgiz, 1962 (Russian edition).
98. Mikhlin, S. G., *Integral Equations and Their Applications to Some Problems of Mechanics, Mathematics, Physics and Engineering*, 2nd edition, Moscow, Gostekhizdat, 1949 (Russian edition).
99. Mikhlin, S. G., *Minimum Problem for a Quadratic Functional*, Moscow, Gostekhizdat, 1952 (Russian edition).
100. Mikhlin, S. G., *Lectures on Integral Equations*, Moscow, Fizmatgiz, 1959 (Russian edition).

101. Mikhlin, S. G., *Multidimensional Singular Integrals*, Moscow, Fizmatgiz, 1962 (Russian edition).
102. Mikhlin, S. G., *Numerical Realization of Variational Methods*, Moscow, Nauka, 1966 (Russian edition).
103. Mikhlin, S. G., *Variational Methods in Mathematical Physics*, 2nd edition, Moscow, Nauka, 1970 (Russian edition).
104. Mitropolsky, Yu. A., *Nonstationary Processes in Nonlinear Oscillatory Systems*, Kiev, 1955 (Russian edition). The English translation: *Nonstationary Processes in Nonlinear Oscillatory Systems*, USA, Air Technical Intelligence Translations, ATIC-270579, F-TS-9085.
105. Morse, P. M., Feshbach, H., *Methods of Theoretical Physics*, V. 1. New York, McGraw-Hill, 1953.
106. Muskhelishvili, N. I., *Singular Integral Equations*, 3rd edition, Moscow, Nauka, 1968 (Russian edition). The English translation: *Singular Integral Equations; Boundary Problems of Function Theory and Their Applications to Mathematical Physics*, Groningen, 1953.
107. Myškis, A. D., *Introductory Mathematics for Engineers. Lectures in Higher Mathematics*, Moscow, Mir Publishers, 1972 (English edition).
108. Panovko, Ya. G., *Elements of Applied Theory of Elastic Vibrations*, Moscow, Mir Publishers, 1971 (English edition).
109. Panovko, Ya. G., Gubanova, I. I., *Stability and Vibrations of Elastic Systems*, Moscow, Nauka, 1967 (Russian edition).
110. Petrovsky, I. G., *Lectures on the Theory of Integral Equations*, Moscow, Mir Publishers, 1971 (English edition).
111. Pliss, V. A., *Some Problems of the Theory of Stability in the Large*, Leningrad, LGU Publishing House, 1958 (Russian edition).
112. Prager, W., *Einführung in die Kontinuumsmechanik*, Basel-Stuttgart, Birkhäuser, 1961.
113. Rashevsky, P. K., *Ricmannian Geometry and Tensor Analysis*, Moscow, Nauka, 1964 (Russian edition).
114. Reinfeld, N. V., Vogel, W. R., *Mathematical Programming*, Englewood Cliffs, N. J., Prentice-Hall, Inc., 1958.
115. Romakin, M. I., *Elements of Linear Algebra and Linear Programming*, Moscow, Vysshaya Shkola, 1963 (Russian edition).
116. Schouten, J. A., *Tensor Analysis for Physicists*, Oxford, Clarendon Press, 1951.
117. Sedov, L. I., *An Introduction to Mechanics of Continuous Medium*, Moscow, Fizmatgiz, 1962 (Russian edition).
118. Shilov, G. E., *An Introduction to the Theory of Linear Spaces*, 2nd edition, Moscow, Nauka, 1956 (Russian edition). The English translation: *An Introduction to the Theory of Linear Spaces*, Englewood Cliffs, N. J., Prentice-Hall, Inc., 1961.
119. Shtokalo, I. Z., *Linear Differential Equations with Variable Coefficients*, Kiev, AN USSR Publishing House, 1960 (Russian edition).
120. Smirnov, V. I., *A Course of Higher Mathematics*, V. 3, Part 2, 8th edition, Moscow, Nauka, 1969 (Russian edition). The English translation: *A Course of Higher Mathematics*, Reading, Mass., 1961.
121. Smirnov, V. I., *A Course of Higher Mathematics*, V. 4, Moscow, Fizmatgiz, 1958 (Russian edition). The English translation: *A Course of Higher Mathematics*, Reading, Mass., 1961.
122. Smirnov, V. I., Krylov, V. I., Kantorovich, L. V., *The Calculus of Variations*, Leningrad, KUBUCh, 1933 (Russian edition).
123. Stoilow, S., *Teoria functiilor de o variabilă complexă*, Bucharest, Ep. Acad. Republ. Popul. Române, V. 1, 1958, V. 2, 1958.
124. Stoker, J. J., *Nonlinear Vibrations in Mechanical and Electrical Systems*, New York-London, Interscience Publ., Inc., 1954.

125. Strelkov, S. P., *An Introduction to the Theory of Vibrations*, 2nd edition, Moscow, Fizmatgiz, 1964 (Russian edition).
126. Sveshnikov, A. G., Tikhonov, A. N., *The Theory of Functions of a Complex Variable*, Moscow, Mir Publishers, 1971 (English edition).
127. Teodorchik, K. F., *Auto-Oscillation Systems*, Moscow, Gostekhizdat, 1952 (Russian edition).
128. Timoshenko, S. P., *Vibration Problems in Engineering*, New York-Toronto-London, Van Nostrand Comp., 1955.
129. Tranter, C. J., *Integral Transformations in Mathematical Physics*, London, Methuen, 1966.
130. Tricomi, F. G., *Integral Equations*, New York, Interscience Publ., Inc., 1957.
131. Tslaf, L. Ya., *Calculus of Variations and Integral Equations*, Moscow, Fizmatgiz, 1966 (Russian edition).
132. Tsypkin, Ya. Z., *The Theory of Linear Impulse Systems*, Moscow, Fizmatgiz, 1963 (Russian edition).
133. Vainberg, M. M., *Variational Methods of Investigation of Nonlinear Operators*, Moscow, Gostekhizdat, 1956 (Russian edition).
134. Vainberg, M. M., Trenogin, V. A., *The Theory of Branching of Solutions of Nonlinear Equations*, Moscow, Nauka, 1969 (Russian edition).
135. Vekua, N. P., *Systems of Singular Integral Equations and Some Boundary-Value Problems*, Moscow, Gostekhizdat, 1950 (Russian edition).
136. Ventsel, E. S., *Elements of Games Theory*, 2nd edition, Moscow, Fizmatgiz, 1961 (Russian edition).
137. Wasow, W. R., *Asymptotic Expansions for Ordinary Differential Equations*, New York, Interscience Publ., Inc., 1965.
138. Yudin, D. B., Golstein, E. G., *Linear Programming*, Moscow, Fizmatgiz, 1963 (Russian edition).
139. Zabreiko, P. P., Koshelev, A. I., Krasnoselsky, M. A., Mikhlin, S. G., Rakovshchik, L. S., Stetsenko, V. Ya., *Integral Equations*. Reference Book, Moscow, Nauka, 1968 (Russian edition).
140. Zeldovich, Ya. B., Myškis, A. D., *Elements of Applied Mathematics*, 3rd edition, Moscow, Nauka, 1972 (Russian edition).
141. Zukhovitsky, S. I., Avdeeva, L. I., *Linear and Convex Programming*, 2nd edition, Moscow, Nauka, 1967 (Russian edition).

Name Index

Abel, N. H. 545, 609
Airy, G. B. 644 ff
Aizerman, M. A. 718, 770
Akhiezer, N. I. 770
Andronov, A. A. 671, 755, 770
Angot, A. 770
Avdeeva, L. I. 775

Babakov, I. M. 770
Banach, S. 324
Barbashin, E. A. 770
Bellman, R. 770
Beltrami, E. 307
Berezin, I. S. 770
Bernoulli, Jakob 318
Bernoulli, Johann 318
Bertrand, J. L. F. 573
Bessel, F. W. 160, 184, 462, 542
Birkhoff, G. D. 680
Blackwell, D. B. 770
Bliss, G. A. 770
Bogolyubov, N. N. 508, 755, 762, 770
Bohl, P. 251, 629
Bohr, H. 627, 770
Bolyai, J. 419
Bolza, O. 352
Bolzano, B. 89, 119, 174, 395
Borisenko, A. I. 770
Brillouin, L. 643
Brouwer, L. E. J. 251
Bubnov, I. G. 643
Bulgakov, B. V. 770
Bunyakovsky, V. Ya. 241

Carslaw, H. S. 770
Cauchy, A. L. 38, 80, 104, 105, 108, 200, 241, 289
Cayley, A. 227
Chaplygin, S. A. 349
Chebotarev, N. G. 770
Chebyshev, P. L. 507, 551
Chesari, L. 771
Chetaev, N. G. 691, 771
Christoffel, E. R. 138, 304
Coddington, E. A. 771
Collar, A. R. 771
Copson, E. T. 771
Cotes, R. 507
Coulomb, C. A. 29
Courant, R. 215
Cunningham, W. I. 771

Davenport, B. Jr. 771
De Bruijn, N. G. 771
Demidovich, B. P. 771
Den Gartog, J. P. 771
Dirichlet, P. G. L. 467
Ditkin, V. A. 771
Doetsch, G. 771
Dorodnitsyn, A. A. 751
Duffing, G. 735, 737, 742
Duncan, W. J. 771

Einstein, A. 280
Eisenhart, L. P. 771
Elsgolts, L. E. 771, 773
Erdélyi, A. 771
Erugin, N. P. 771
Euler, L. 29, 82, 93, 136, 317, 333, 343, 484, 667

Faddeev, D. K. 771
Faddeeva, V. N. 771
Fermat, P. 421
Feshbach, H. 774

Filchakov, P. F. 771
Fisher, E. S. 546
Floquet, G. 611
Fomin, S. V. 772
Forsythe, G. E. 771
Fourier, J. B. J. 149, 170, 195
Frazer, R. A. 771
Fredholm, E. I. 486, 488
Fröman, N. 771
Fröman, P. 771
Fuks, B. A. 771

Gakhov, F. D. 771
Galerkin, B. G. 463
Gantmacher, F. R. 770, 771
Gass, S. 772
Gauss, K. F. 16, 238, 308, 419, 505
Gegenbauer, L. 551
Gelfand, I. M. 772
Germain, P. 772
Girshick, M. A. 770
Golstein, E. G. 775
Gorelik, G. S. 772
Gram, J. P. 477
Green, G. 487
Gubanova, I. I. 774
Guter, R. S. 6

Hadamar, J. S. 547
Hadley, G. 772
Hale, Jack K. 772
Halmos, P. R. 772
Hamilton, R. W. 16, 227, 402, 403, 416, 423, 426, 429
Hammerstein, A. 593
Hankel, H. 160, 201
Hayashi, Ch. 772
Heading, J. 772
Heaviside, O. 162, 183, 674
Hermite, C. 209, 488, 551
Hilbert, D. 198, 324, 486, 490, 532, 577
Hill, W. 614
Hooke, R. 433
Hopf, E. 564
Hurwitz, A. 127f

Ince, E. L. 772

Jakobi, K. G. J. 219, 245, 378, 416
Jeager, J. C. 770
Jordan, C. 110, 224, 229

Kantorovich, L. V. 257, 481, 772, 774
Karacharov, K. A. 772
Karpelevich, F. I. 772
Kauderer, H. 772
Kellog, O. D. 530
Khaikin, S. E. 770
Kharkevich, A. A. 772
Khinchin, A. Ya. 772
Kochin, N. E. 772
Kontorovich, M. I. 772
Kopachevsky, N. D. 6
Koppenfels, W. 772
Koshelev, A. I. 775
Kramers, H. A. 643
Krasnoselsky, M. A. 6, 772, 775
Krasnov, M. L. 772
Krasovsky, N. N. 691, 772
Krein, M. G. 525, 623, 771
Krelle, W. 773
Kroneker, L. 283
Krylov, A. N. 252
Krylov, N. M. 456, 508, 755, 762
Krylov, V. I. 772, 773, 774
Kurosh, A. G. 773
Kuznetsov, P. I. 771
Künzi, H. P. 773

Lagrange, J. L. 213, 317, 345, 349, 423, 705
Laguerre, E. N. 551
Lamé, G. 305
Lance, G. N. 773
Laplace, P. S. 21, 29, 162, 307, 563
La Salle, J. 773
Lavrentyev, M. A. 773
Laurent, P. M. H. 80, 101
Lebesgue, H. 489
Lefschetz, S. 773
Legendre, A. M. 159, 373, 551
Leibniz, G. W. 79, 163, 318
Letov, A. M. 773
Levi-Civita, T. 280
Levin, V. I. 771
Levinson, N. 771
Levitan, B. M. 773
L'Hospital, G. G. A. 102, 318
Lichtenstein, L. 592
Lie, M. S. 413
Lindelöf, E. L. 74
Liouville, J. 119, 609, 679
Lobachevsky, N. I. 419, 420
Lorentz, H. 297
Lovitt, W. V. 773
Luce, R. D. 773
Lunts, G. L. 6, 773
Lurye, A. I. 713, 773

Lyapunov, A. M. 445, 592, 613, 623, 668, 685
Lyusternik, L. A. 773

Makarenko, G. I. 772
Malkin, I. G. 773
Markushevich, A. I. 773
Maron, I. A. 771
Mayer, A. 352
Mathieu, E. L. 614
McConnell 773
McKinsley, J. C. C. 773
McLachlan, N. W. 773
Melentyev, P. I. 773
Mellin, R. H. 198
Mikhailov, A. V. 130
Mikhlin, S. G. 773, 774, 775
Milne, E. 567
Mitropolsky, Yu. A. 755, 770, 774
Mladov, A. G. 6
Möbius, A. F. 44
Moler, C. B. 771
Morera, A. F. 87
Morse, P. M. 774
Muskhelishvili, N. I. 774
Myškis, A. D. 774, 775

Nazarov, N. N. 604
Nekrasov, A. I. 604
Noether, A. E. 412
Neumann, J. 270, 272
Neumann, K. G. 511
Newton, I. 32, 79, 318, 507

Ostrogradsky, M. V. 16, 343, 609

Panovko, Ya. G. 774
Parseval, M. A. 171
Perov, A. I. 772
Petrov, G. I. 463
Petrovsky, I. G. 774
Pilyutik, A. G. 772
Plateau, J. A. F. 319
Pliss, V. A. 774
Poincáre, H. 573, 623, 645, 660, 669, 670
Poisson, S. D. 32, 116, 190, 404, 442
Pol, B. van der 747, 755
Pontryagin, L. S. 671
Popov, V. N. 715
Povolotsky, A. I. 772
Prager, W. 774
Prudnikov, A. P. 771
Puiseux, V. A. 97, 101

Raiffa, H. 773
Rakovshchik, P. K. 775
Rashevsky, P. K. 774
Rayleigh, J. 215
Reinfeld, N. V. 774
Ricci, C. G. 280
Riemann, G. F. B. 38, 43, 60, 71, 311, 489, 577
Riesz, F. 524, 546
Ritz, W. 453, 455, 465, 472, 475
Romakin, M. I. 774
Root, W. L. 771
Rouché, E. 122
Routh, E. 128
Rutitsky, Ya. B. 772

Sadovsky, L. E. 772
Schauder, J. 596
Schmidt, E. 532
Schouten, J. 774
Schwarz, H. A. 138, 241
Sedov, L. I. 774
Segre, C. 230
Shabat, B. V. 771, 773
Shilov, G. E. 772, 774
Shtokalo, I. Z. 774
Smirnov, V. I. 774
Sokhotsky, Yu. V. 103
Stallman, F. 772
Stetsenko, V. Ya. 772, 775
Stieltjes, T. J. 489
Stirling, J. 155
Stoilov, S. 774
Stoker, J. J. 774
Stokes, G. G. 16, 146
Strelkov, S. P. 775
Sturm, J. C. F. 637
Sveshnikov, A. G. 775
Sylvester, J. J. 218, 220

Tarapov, I. E. 770
Taylor, B. 85
Teodorchik, K. F. 775
Tikhonov, A. N. 547, 775
Timoshenko, S. P. 775
Tranter, C. J. 775
Trefftz, E. 471
Trenogin, V. A. 775
Tricomi, F. G. 775
Tslaf, L. Ya. 775
Tsypkin, Ya. Z. 775
Tyuptsov, A. D. 6

Uryson, P. S. 589

Vainberg, M. M. 775
Vainikko, G. M. 772
Vandermonde, A. T. 507
Vekua, N. P. 775
Ventsel, E. S. 775
Vitt, A. A. 770
Vogel, W. R. 774
Volterra, V. 486, 488

Wasow, W. 775
Weierstrass, K. T. W. 89, 119, 174, 367, 386, 395
Wentzel, G. 643
Wiener, N. 564
Wronski, J. M. 608

Young, T. 433, 442
Yudin, P. B. 775

Zabreiko, P. P. 772, 775
Zeldovich, Ya. B. 775
Zermelo, E. F. F. 270
Zhidkov, N. P. 770
Zhukovsky, N. E. 54
Zukhovitsky, S. I. 775

Subject Index

Abel's integral equation 545
Action 424, 427, 438
  variable 732
Activity-analysis problems 259
Acyclicity 28
Adjoint
  (associated) integral equation 493
  mapping 209
  system of differential equations 607
Affine tensor 293
Airy's
  equation 644
  function 645
Algebraic topology 667, 677
"Almost everywhere" 680
Almost periodic function 627
Alpha-limit set 648
  spectrum and amplitudes of 628
Amplification factor (gain coefficient) 183, 721
Amplitude of an elliptic integral 142
Amplitude-frequency characteristic 721
Analytic
  continuation 91
  function 36
    complete 91
    entire 87
    implicit 96
    meromorphic 192
    of a real argument 94f
    of several arguments 95f
    singular point of 37
Angular variable 732
Annulus of convergence 81
Anti-Hermitian (skew-Hermitian) kernel 533
Antisymmetric (skew-symmetric) tensor 287
Aperiodic motion 720
Applicable (isometric) surfaces 311
A-point of a function 86
Argument principle 121
Associated tensor 295
Asymptotic expansion 144
  of some integrals 149, 153
  properties of 146ff
Asymptotical stability 685
  in the large (global) 691
Automatic control system 713
  absolutely stable 714
Autonomous system 183, 221, 428
  of differential equations 647
Auto-oscillation 658
Axioms
  of linear space 204
  of metric space 48
  of norm 322
Axis of vortex (whirl) 64

Banach space 324
Bar
  equation of longitudinal vibrations of 434
  equation of transversal vibrations of 435
Barrier potential 449
Barycentric coordinates 262
Basis (bases) 204
  biorthogonal 211
  Euclidean 205
  reciprocal 293
Beating 722
Bending of a surface 310
Bessel's
  equation 184, 462
  functions (of the 1st, 2nd and 3rd kind) 160f
  inequality 542

Best conditioned matrix 243
Bifurcation
  equation 602
  point 589, 603
Biharmonic
  equation 437, 469
  operator 437
Birkhoff's theorem 680
Block (submatrix) of a matrix 207
Bohl-Brouwer (fixed point) theorem 251
Bolzano-Weierstrass theorem 89, 119 174, 395
Bolza's problem 352
Bordering 716
Boundary
  condition(s), 67, 333
    of the 1st, 2nd and 3rd kind 434
    natural 354, 362, 391, 393
  layer 753
  point
    multiple 72
    simple 72
  value problem 67, 184
    exterior 67
    interior 67
Brachistochrone problem 318f
Branch of a function 51
Branch point 51
  algebraic 52
  logarithmic 53
  of finite order 52
Broken extremal 365ff
Bubnov-Galerkin method 463
Buckling of a bar 447ff
Bulk (compressibility) modulus 442

Calculus of variations 244, 317ff
Canonical form
  of Euler's equations 403
  of the equation of a second-order surface 217
  of a quadratic form 216
  system of differential equations 403
  transformation 405
    generating function of 406
  variables 402
    conjugate 402
Cartesian
  basis 205
  tensor 284
Catenary 338
  problem 447
Catenoid 338
Cauchy-Bunyakovsky-Schwarz inequality 241
Cauchy-Riemann equation 38
Cauchy's
  integral formula 105
  integral theorem 80
  kernel 572
  principal value of a singular integral 108, 200
  residue theorem 104ff
Central field of extremals 378
Centre 652
Chaplygin's problem 349ff
Characteristic
  amplitude-frequency 721
  equation 211
  function of a set 681
  matrix 185
  multiplier 612
  of a controller 713
  of a restoring force 728
  polynomial 181
  quasi-polynomial 187
  value (number) of an integral equation 493
    extremal properties of 536ff
Chebyshev's
  integration method 507
  polynomials 551
Christoffel's symbol
  of the 1st kind 304
  of the 2nd kind 303
Circle of convergence 86
Circle preserving property 45
Circulation
  of a vector field 16
  of a tensor field 301
Classical canonical matrix 229
Collocation method 483
Combination frequency 758
Combinatorial topology 677
Commuting matrices 235, 237f
Compactness 395
Compatibility condition 348
Competition of strategies (games theory) 270
Complementary modulus of an elliptic integral 142
Complete analytic function 91
Complete elliptic integrals 142
Completely continuous operator 524
Completeness 323, 512
Completion 323
Complexification 541
Complex potential 62
Component of a tensor 281

Composition (inner multiplication) of tensor 287
Computing eigenvalues 244ff
Condition number of a matrix 242
Conditional
  extremum 213, 348ff
  stability 686
  stationary value 427
Condition (Weierstrass'), sufficient for strong extremum 385ff
Cone in a functional space 525
Configuration space of a system 451
Conformal
  image 41
  mappings of the 1st and 2nd kind 41
Conjugate
  (dual) mapping 208
  harmonic functions 39
  matrix 205
  points 375, 377
Connectivity number 26
Conservative system 428
Constraint (subsidiary condition) 213, 257, 348
  bilateral (two-sided) 363
  differential (nonholonomic) 349
  geometric (or holonomic or finite) 348
  integral 344
  non-restricting (or one-sided or unilateral) 257, 363
Contact transformation 405, 408
  generating function of 308
Continued fraction 504
Continuous (Lie) group of transformations 411, 413
  infinitesimal generator of 411
Contraction mapping 513
  principle 513
Contraction of a tensor 286
Contravariant
  index 292
  tensor 292
Control
  direct 713
  indirect 713
Control action (or steering signal or controlling influence) 713
Controlled object 713
Convex
  cone 261, 265
  function 278
  hull of a set 261
  polyhedron 261
  programming 278
  set of points 260
  set of vectors 261
Convolution
  of functions 168
    Laplace transform of 169
  of kernels 510
Coordinate functions 455
Cosinus amplitudinis (Jacobi's "cn"-function) 143
Cost function 257
Coulomb's law 29
Coupling matrix 712
Courant's theorem 215f, 399
Covariant
  derivative of a field 305f
  index 292
  tensor 292
Critical
  cases of stability 698ff
  force 448
Curvature 315
  tensor 315
Cut 50
Cycle 648, 658
  half-stable 658
  isolated 658
  limit 658
  stable 658
  unstable 658
Cycloid 337

Damped oscillations 720
Damping
  decrement 720
  factor 720
Degenerate
  kernel 491
  node 653
Degree of stability 693
Del (nabla) 16
  repeated operations with 20f
Delay theorem 167
Deleted neighbourhood 82
Delta amplitudinis (Jacobi's "dn"-function) 143
Density of Hamiltonian function 431
Derivative of a tensor field 300
Determining equation 736
Deviation of functions 322f
Dido's problem 317
Diet problem 258
Difference-differential equation 186
Difference equation 185
Differential equation(s) 606ff
  abjoint 607
  self-adjoint 608
Dipole 30
  polarization (dipole moment) of 30

Direct
control 713
methods in variational problems 454
sum (decomposition) 206
orthogonal 209
properties of 206
summands 208
Dirichlet's integral 467
Discontinuous systems 672ff
Discriminant of a polynomial equation 131
Discrete programming 276
Discrete-time system 765
Dissipative
force 711
system 443ff
Distance between elements of a functional space 323
Distance function (metric) 418
Divergence 15, 306
expression 426, 438
of a tensor field 301
Domain of influence (of attraction) of a limit cycle 658
Double layer 66
Double point of a boundary 72
Doubly
connected domain 26
periodic function 143
Dual (conjugate) mapping 208
Duffing's equation 735, 737, 742
Dummy (umbral) index 282
Dyad 286
Dynamic programming 277
Dynamic system 680

Eigenfunction 388, 494
Eigenvalue 210, 388
extremal property of 213ff, 387ff
Eigenvector 210, 388
Element (vector) of a linear space 204
Elementary divisor of a matrix 230
Elimination scheme (Gauss' method) 238ff
Ellipsoid 217
of inertia 289
Elliptic integrals 141ff
in Legendre's normal (standard) form 142
Elliptic (Jacobi's) functions 143
Empty (or void or null) set 263
Energy density of a field 439
Energy-flux density 440
Equation
Hamilton-Jacobi 416
Laplace's 29
of oscillations of an elastic medium 443
of vibrations of a string 431
Poisson's 33
Ergodic
system 681
theorem 680
Error function 173
Euclidean
basis 205
space 204
Euclid's parallel postulate 419
Euler-Lagrange equations 438
Euler-Ostrogradsky equation 343
Euler's
characteristic 667
constant 136
equation(s) 333ff, 339, 340, 343
canonical form of 403
identity 341
invariance of 336
vector form of 340
method (in the calculus of variations) 484
theorem on homogeneous functions 428
transformation 93
Everywhere dense trajectory 684
Expansion of a symmetric kernel with respect to its eigenfunctions 528
Extended complex plane 43
Extension of a function 92
Extremal 335
reflection of 365
refraction of 366
with corner points 365ff
with moving boundaries 353f, 357ff
Extremum of a functional 324
conditional 348ff
Legendre's necessary conditions for 373
necessary conditions for 331ff
strong 324
sufficient conditions for 371f, 378
weak 324

Factorization 565
Fast time 752
Feedback signal 713
Fermat's principle 422
Field
circulation-free (or non-circulatory) 24
irrotational 24
of extremals 378
of a tensor 300

potential 23
solenoidal 26
variables 438
vector potential of 26
Finite-dimensional operator 523
First integral 336, 404
Fixed point of a mapping 42, 251, 513
Floquet's theorem 611
Flow 648
Flux
of a vector field 16
of a tensor field 301
Focal (spiral) point 652
Forced oscillations 734ff, 743ff
Form of several variables 216
linear 216
metric 295
quadratic 216
reduction to canonical form of 217
Formula
Ostrogradsky-Gauss 16, 302
Stokes's 16
Fourier's
cosine transformation 197f
sine transformation 197f, 202
transformation 149, 195ff
inversion formula of 195f, 200
n-dimensional 200
of a function of a bounded order of growth 195ff
two-dimensional 200
transform of a function 150, 170, 195
type of integral 149
Fredholm's
determinant 493
equation of the 1st kind 488, 548ff
with difference kernel 488
with symmetric kernel 488, 545ff
equation of the 2nd kind 488
first alternative 498
first theorem 493
fourth theorem 498
second alternative 498
second theorem 495
third theorem 496
Free index 282
Frequency characteristic 183
Function
almost periodic 627
analytic 36
at the point $z=\infty$
convex 278
entire 87
finite (of a compact support) 32
harmonic 29, 39
index of growth of 195
meromorphic 132
multivalent (many-sheeted) 40
normalized 389
objective (or cost or utility) 257
of a matrix 233ff
of an exponential rate of growth 163
sourcewise representable with the aid of a kernel 532
square-summable 322
summable 322
transfer 182
univalent (one-sheeted) 40
variation of 326
Zhukovsky's 54
Functional 320
bilinear 370
bounded linear 329
norm of 329
conditional extremum of 348ff
continuous 330
domain of definition of 321
extremum of 324
first variation of 327, 328
involving derivatives of higher order 338f
linearization of 328
multilinear 370
n-th variation of 369
of functions of several variables 341ff
of several functions 339ff
quadratic 370, 374ff
bounded 370
norm of 370
stationary value of 335
Functional analysis 322
Functional space 322f
Banach 324
element (generalized vector) of 323
Hilbert 322
normed linear 322
of continuous functions 322
of continuously differentiable functions 322
Fundamental
coefficient of the 1st order of a surface 308
(covariant) metric tensor 308
(first) quadratic form of a surface 308
matrix of a system of differential equations 609
metric tensor 294
system of solutions 607
theorem of algebra 123

Gain coefficient (or amplification factor) 183
Galerkin (Bubnov-Galerkin) method 463
Game(s)
  a play of 270
  theory 270
  value of 275
Gamma function 92
Gauss'
  integration method 505
  method (elimination scheme) 238ff
Gegenbauer's polynomials 551
General affine tensor 293
Generalized
  coordinates 221, 427
  distance 418
  momentum 428
  momentum density (canonical momentum density) 439
  solution 368
  vector (element of a functional space) 204, 490
Generating function
  of a canonical transformation 406
  of a contact transformation 408
  of a functional sequence 550
Geodesic 382
Global stability 691
Gradient 15, 306
  method 476
Grammian determinant 477
Green's function 486
Gyroscopic force 711

Half-plane 47
Hamilton-Cayley theorem 227
Hamilton-Jacobi equation 416
Hamilton's
  action 424, 438
  form of Euler's equations 403
  function 402
    density of 431
  operator 16
  principle 423ff, 426ff, 429ff
    for continuous media 429ff
Hammerstein equation 593
Hankel's
  function of the 1st kind (Bessel's function of the 3rd kind) 160
  transformation 201
Hard characteristic of a restoring force 728
Harmonic functions 29, 39
Heat conduction equation 188
Heaviside (unit step) function 183, 674
Hermitian
  kernel 488
  operator 541
  (or self-adjoint) matrix 209
  quadratic form 217
Hermite's polynomials 551
Hilbert's
  kernel 572
  space 324, 490ff
  transformation 198ff
Hilbert-Schmidt theorem 532
Hill's equation 614
Holonomic constraint 348
Homeomorphism 43, 44
Hooke's law 433
Hurwitz
  criterion 128
  determinant 128
  (stable) polynomial 127
  theorem 128
Hyperbolic point (saddle-point) 157
Hyperboloid 217
Hyperplane 260, 333
Hysteresis
  phenomenon 688
  loop 688

Ideal (perfect) liquid 61f
Identity (or unit) operator 522
Ill-conditioned system of equations 242
Image (under a mapping) 210
Improper point (or point at infinity) of the complex plane 43
Improperly posed problem 547
Impulse function 195
Incomplete elliptic integrals 142
Index
  contravariant 292
  covariant 292
  dummy (umbral) 282
  free 282
  of a curve relative to a vector field 661
  of a quadratic form 218
  of a Riemann-Hilbert problem 578
  of a singular point of a plane vector field 662
  of growth of a function 195
Indirect control 713
Inertia
  ellipsoid of 289
  index of 218
  law of a quadratic form 218
  moments of 289
  principal axes of 289
  products of 289
  tensor 289
  zone 674

Infinite product 134
Influence function 183, 486
Initial function 162
Inner product of tensors 287
Input 182
Input-output system 182
Integral
  constraint 344
  equation 187, 486ff
    Fredholm's 488
    homogeneous adjoint (associated) 494
    resolvent (reciprocal) kernel of 493
    singular 489, 569ff
    with positive kernel 524f
    Volterra's 488, 519ff
  improper 78
  invariant 678
    density of 678
  manifold 733
  of a complex function 78f
  of an analytic function 79f
  of a tensor field 301f
  of the potential type 518
  operator 490
    anti-Hermitian 542
    Hermitian 541
  singular 108, 569ff
    principal value of 108, 200
  transformation 192, 197ff
    on a finite interval 201ff
Integro-differential equation 187
Interchanging indices in a tensor 287
Intersection (or product or meet) of sets 466
Interval of convergence 87
Intrinsic
  (absolute) properties of a surface 311
  geometry of a surface 311
Invariance of Euler's equations 336
Invariant
  circle 45
  measure 679
  subspace of a mapping 206
Inverse transform (preimage) 162
Inversion 45f
  formula
    for Hankel's transformation 201
    for Hilbert's transformation 200
    for Fourier's transformation 195f
    for Laplace's transformation 170
    for Mellin's transformation 198
Involution 45
Irreversibility of time 299
Irrotational field 24
Isochronism 721
Isolated singular point 100
Isometric (applicable) surfaces 311
Isomorphism 205
Isoperimetric
  (integral) condition 345
  problem 344ff
Isotropy 420, 436
Iterated kernel 510

Jacobi's
  elliptic functions 143
  equation 378
  method
    for computing eigenvalues 245
    for reducing a quadratic form to canonical form 219
  sufficient conditions for extremum of a functional 378
Jordan's
  block (submatrix) 221
  lemma 110
  normal (classical canonical) form of a matrix 229

Kellog's method 530
Kernel 197, 488
  antisymmetric (or skew-symmetric or anti-Hermitian) 533
  Cauchy's 572
  characteristic value of 499
  degenerate 491
  difference 488
  Hermitian 488
  Hilbert's 572
  iterated 510
  of an integral transformation 197
  perturbed 515ff
  positive 524
  positive-definite 537
  positive-semidefinite 537
  resolvent (or reciprocal) 493
  symmetric 488
    expression in eigenfunctions of 528
  with polar singularity 489
Kinetic energy 423
Kinetic potential (Lagrangian function) 423
Klein-Gordon equation 432
Kroneker delta 283
Krylov-Bogolyubov method 762

Lagrange's
equations 438
function (kinetic potential) 423
density of 431
multipliers 213f, 345, 349
problem 349
theorem 705
Laguerre's polynomials 551
Lamé's coefficients 305
Laplace-Beltrami operator 307
Laplace's
equation 29, 467
inverse transformation 170
operator (Laplacian) 21, 308
transformation 162
applications of 180ff
discrete 193
inversion formula for 170
transform(s) 162
of a convolution 168
of the delta function 165
of a derivative 167f
of an exponential function 164
of an integral 168
properties of 166ff
table of 166
two-sided transformation 563
Laurent-Puiseux series 101
Laurent's series 80ff
principal and regular parts of 81
Law
of conservation of energy 428, 431
of conservation of momentum 428
of inertia for a quadratic form 218
Least square method 479
Legendre's
necessary conditions 373
polynomials 159, 506, 551
Leibniz rule 163
Level line 62
L'Hospital's rule 102
Libration 730
Lie group 413
Light cone 299
Limit cycle (stable or unstable or half-stable) 658
Lindelöf's theorem 74
Linear algebra 205
Linear functional 321
bounded 329
norm of 329
Linearization of a functional 328
Linear mapping 42, 205
"onto" and "into" 205
of a real space 230ff
Linear operations on tensors 285
Linear operator 209, 490
anti-Hermitian 542
bounded 491
norm of 491
Hermitian 541
Linear oscillator 424
Linear programming 256ff
application to matrix games of 270ff
basic problem of 256f
basic solution of 269
examples of 258f
feasible solution of 269
geometric interpretation of 262ff
optimal solution of 259, 269
solution of 269
standard form of 264f
terminology of 269
Linear space 204
Liouville's theorems 119, 679
Lobachevskian
geometry 419
plane 420
Logarithmic
decrement 720
potential 33
residue 120
Logarithm of a matrix 235
Lorentz
plane 297
rotation 298
space 297
transformation 297
Lowering indices in a tensor 295
Lyapunov-Lichtenstein equation 592
Lyapunov-Poincaré theorem 623
Lyapunov's
definitions of stability and instability 685
first method 692
function 689
positive definite 690
second method 692
stability of an equilibrium state 455
theorem on stability 689

Magnification (or amplification) factor (or gain coefficient) 183, 721
Mapping 39
adjoint 209
analytic 40, 88ff
conformal, of the 1st and 2nd kind 41
conjugate (dual) 208
continuous 43
fractional-linear 44

involutary 45
linear 42, 205
multivalent (or many-sheeted or many-to-one) 40
norm of 241
null space of 209
of a real space 230ff, 236f
of a space into itself 205
one-valued 40
self-adjoint 209, 212f
univalent (or one-to-one or one-sheeted) 40
Mathieu's
equation 614
functions 619
Matrix
characteristic 185
classical canonical 229
conjugate 205
differential equation 608
elementary divisor of 230
equation 185
function of 233ff
Jordan's form of 229
logarithm of 235
monodromy 612
norm of 241
of a mapping 206
of a quadratic form 216
orthogonal 213
oscillatory 252
partition into blocks of 207
payoff 270
polynomial (or λ-matrix) 227
principal minors of 219
quasi-diagonal 207
Segre's characteristic of 230
self-adjoint (Hermitian) 209
simple classical 225
symmetric 209
trace of 287
transfer 185
transpose of 205
triangular (semi-diagonal) 240
unitary 213
with nonnegative elements 250ff
Maximum-modulus principle 91
Mayer's problem 352
Mean curvature 319
Mean ergodic theorem 680
Mean-square deviation of functions 323
Mean-value theorem 117
Mellin's transformation 198
Membrane 432
equation of vibrations of 432
Merit criterion 257
Meromorphic function 132
Method
Gauss' 238ff
of averaging 758
of harmonic balance 643
of Jacobi 219, 245f
of Kantorovich 481
of Krylov 252f
of Lagrange's multipliers 213f, 345
of Petrov 463
of steepest descent 158, 243
operational 180
saddle-point 156ff
small-parameter 253f, 592f
Metric 418
form 295
space 418, 511
tensor 294
Mikhailov's criterion 130
Milne's equation 561
Minimal surface 319
Minimizing sequence 393
Minimum-maximum principle (Courant's minimax theorem) 215f
Minimum of a functional
basic condition for 396ff
existence of 393ff
Minimum-surface-of-revolution problem 320, 379ff
Mixed
strategy 271
tensor 292
Mixing property 681
Möbius transformation (fractional-linear mapping) 44
Modulus (longitudinal) of elasticity (or Young's modulus) 433, 442
Modulus of an elliptic integral 142
Moment
of a function 554
of inertia 289
Momentum density 431
of a field 441
Monofrequent vibration 707
Move (in a play of a game) 271
Morera's theorem 87
Multiple boundary point 70,72
Multiple root 86
Multipliers of a periodic system of differential equation 612
Multiply connected domain 26
Multistage decision process (dynamic programming) 277

Nabla (del) 16
repeated operation with 20f

Natural
boundary conditions 354, 362, 391, 393
frequency 707
mode of vibration 707
trajectory 421
Nazarov-Nekrasov method 604
Neighbourhood of the point $z=\infty$ 43
Necessary conditions for extremum of a functional 331ff
Negative friction 444
Net methods 455
Neumann series 511
Neutral equilibrium 450
Newton-Cotes integration formula 507
Newtonian potential 32
Newton-Leibniz formula 79
Newton's polygon 99, 132
Nodal-focal singular point 666
Nodal point 652
Noether's theorem 412, 413
Non-circulatory field 24
Nonclassical minimax problem 451
Non-degeneracy condition in linear programming 267
Nonlinear
integral equation 587ff
programming 278
Nonoscillating solution of a differential equation 637
Norm
of a bounded linear functional 329
of a linear operator 491
of a mapping 241
of a matrix 241
of an element of a functional space 322
of a quadratic functional 370
of a vector 205
Normal
mode of vibration 706
(principal) coordinates 222
Normalization 215, 247, 248f, 254
coefficient 390
condition 389, 453
of a scalar potential 24
of a vector potential 26
Normalized system of functions 389, 527
Nucleus (kernel) of an integral equation 488
Null space of a mapping 209
Number vector 205

Objective function 257
Omega-limit set 648
Operational calculus 162ff, 180
Operator 16, 209
anti-Hermitian 542
completely continuous (compact) 524
equation 491, 525
finite-dimensional 523
Hermitian 541
linear 491
bounded 491
norm of 491
of the Volterra type 520
Optimal control
problems 422
Optimal strategy 270
Orbital stability 731
Order (rank) of a tensor 284
Order relation 250
Original 162
numerical computation of 177ff
expansion into a series of 174ff
of a Laplace transformation 162
Orthogonal
complement 209
matrix 213
Orthogonality 205
with weight function 551
Orthonormal (Euclidean) basis 205
Oscillating solution of a differential equation 637
Oscillator with clearance 675
Oscillatory matrix 252
Ostrogradsky-Gauss formula 16
for tensors 302
Ostrogradsky-Hamilton integral 431
Outer product of tensors 286
Output 182

Parameter of an elliptic integral of the 3rd kind 142
Parametric
form of a variational problem 341
resonance 620
vibration 726
Parseval's relation 171
Partial differential equation 188
Payoff matrix 270
Period of oscillations 720
Periodic system of differential equations 611ff
Perturbation method (small-parameter method) 74ff, 592ff
Perturbed solution of a differential equation 684

Phase
diagram 674
integral method (WBK-method) 643
space of an autonomous system of differential equations 647
trajectory 648
Plane parallel flow 61
Plate, equation of vibrations of 437
Plateau problem 319
Player (theory of games) 270
Play of a game 270
Poincaré-Bertrand theorem 573
Poincaré's mapping 660, 669
Point at infinity 43
Poisson's
bracket 495
equation 33
integral 115ff
ratio 442
summation formula 196f
Polar singularity 23, 489
integral equations with 521ff
Pole 100ff
multiple 100
of order n 100
simple 100
Polyharmonic excitation 645
Potential 423
complex 62
energy 426
field 23
function 23
kinetic 423
logarithmic 33
Newtonian 32
pit 450
scalar, of a vector field 23
vector, of a vector field 26
Preimage 162
Principal
axes
of inertia 289
of a quadratic form 217
of a second-order surface 217
matrix solution of a homogeneous linear system of differential equations 609
minors of a matrix 219
oscillation (normal mode of vibration) 706
part of a Laurent expansion 81
value
of a divergent improper integral (Cauchy's principal value) 108, 200, 569
of logarithmic function 58
vector 227
Principle
of maximum 91
of minimum of the potential energy 445ff
Probability
of finding the moving point of a trajectory in a set 681
theory 271
vector of a mixed strategy 271
Problem
of Aizerman 718
of Lurye 713f
Product
of inertia 289
of tensors (inner and outer) 286f
Programming
convex 278
discrete 276
dynamic 277
linear 256ff
nonlinear 278
parametric 277
partition 277
quadratic 279
stochastic 278
Proper time 298
Pseudo-Euclidean space 295
Pseudolength 296
Pseudo-Riemannian space 315
Pseudoscalar product 295
Pseudotensor 301
Puiseux series 97
Pulse excitation 626
Pure
deformation 291
rotation 291
strategy 271

Quadratic
form 216
canonical (diagonal) form of 216
definite 217
degenerate 217
diagonal 217f
Hermitian 217
indefinite 295
index of 218
law of inertia of 218
matrix of 216
negative definite 217
positive definite 217
principal axes of 217
rank of 218
reduction
to canonical form of 217
to principal axes of 217

functional 370, 374
signature of 218
programming 279
Qualitative theory of differential equations 692
Quasi-diagonal matrix 207
Quasi-elliptic integral 141
Quasi-periodic function 629
Quasi-polynomial 127, 186
Quasi-static process 688

Radius of convergence 86
Raising indices in a tensor 295
Rank
of a quadratic form 218
(or order) of a tensor 284
Rayleigh's method 215
Real canonical form of a real matrix 233
Reciprocal
algebraic equation 623
bases 293
variational problem 347
Reciprocity principle (in calculus of variations) 347
Reducible system of linear differential equations 613
Reduction
of a quadratic form to a canonical form 217
of a tensor to principal axes 288
Reflection
in a circle 45
transformation 42
Reflection-of-extremal problem 365
Refraction of extremals 366
Regular part of Laurent's expansion 81
Regular point 101
Regularization
of an improperly posed problem 547
of a singular integral equation 575
Relativity of simultaneity of events 298
Relaxation vibrations 750
Residue(s) 101
application to improper integrals of 106ff
at the point $z=\infty$ 120
logarithmic 120
theorem (Cauchy's theorem) 104ff
Resolution of a tensor into symmetric and antisymmetric parts 287f
Resolvent (or reciprocal) kernel 493
Resonance 722
Rest point 648
Restriction
of a function 92
of a mapping 212
Resultant of polynomial equations 13
Riemann-Hilbert problem 577
Riemannian
space 311ff
sphere 43
surface 60
Riesz-Fischer theorem 546
Ritz method 453, 455, 465, 472, 47
Root (or zero) of a function 86
Rotation 15, 40, 307
Lorentz's, through a hyperbolic angle 298
of a tensor field 301
Rotational motion of a pendulum 73
Rouché's theorem 122
Routh-Hurwitz problem 128

Saddle-point method 156ff
Scalar 284
product 204
Schwarz-Christoffel transformation 13
Secular terms 724
Segre characteristic of a classical canonical (Jordan) matrix 230
Self-adjoint
mapping 209, 212f
operator 539
system of linear differential equations 608
Self-sustained oscillations 744
Semi-diagonal (triangular) matrix 240
Separatrix 649
Set
convex 260
empty (or null or void) 263
(sequentially) compact 395
Shauder's theorem 596
Sheet of a Riemannian surface 59
Shift theorem 167
Signature of a quadratic form 218
Signum function 647
Similar matrices 230
Simple
classical (Jordan) matrix 225
layer 65
pendulum 594, 620, 729
point of a boundary 72
Simplex
l-dimensional (l-simplex) 262
method 265ff
Simultaneous reduction of two quadratic forms to canonical form 220ff
Singular
integral 108, 569ff
part of Laurent's expansion 81
point 37, 100

nodal 652
nodal-focal 666
non-isolated 132
saddle 651
saddle-focal 665
spiral (focal) 652
Singularity 37, 100
essential 100
isolated 100
removable 100
Sinus amplitudinis (Jacobi's "sn"-function) 143
Skew-symmetric (antisymmetric) tensor 287
Sliding regime 674
Slow time 751, 755
Slowly varying function 621
Small divisors 645
Small-parameter method 74ff, 592f
Sokhotsky's theorem 103
Soft characteristic of a restoring force 728
Solenoid 27
Solenoidal field 26
Sourcewise representable function 532
Space-like direction 297
Space-time 315
Spectrum
of an almost periodic function 628
of an operator 389
of a symmetric matrix 244f
Spur (or trace) of a matrix 287
Square-summable function 322
Stability 685
absolute 714
asymptotic 685
in the large (global) 691
conditional 686
critical cases of 698ff
degree of 693
in linear (first) approximation 693
in the sense of Lyapunov 685
orbital 721
structural 654, 667f
with respect to constantly acting perturbations 685
Stabilization 716
State space 429
Stationary
point (or a point of rest or an equilibrium solution) of a differential equation 648
value of a functional 335
conditional 427
Steepest descent, method of 158
Step-by-step computation 185
Step function 183
Stereographic projection 43
Stieltjes' integral 489
Stirling's formula 154f
Stochastic programming 278
Stokes' formula 16
Strain tensor 291
Strategy 270f
Stream
function 62
line 62
Stress tensor 442
String 430
equation of vibrations of 431, 487
Stroboscopic method 769
Strong extremum 324
Weierstrass' sufficient condition for 385ff
Sturm's
comparison theorem 638
separation theorem 638
Subharmonic vibration 726, 741ff
Submatrix (or block) of a matrix 207
Subspace 205
invariant under a mapping 206
of a metric space 512
spanned by a system of vectors 208
Summation convention 282
Superharmonic (ultraharmonic) oscillation 741
Superposition principle 190
Sylvester's
law of inertia 218
theorem 220
Symmetric
kernel 488
matrix 209
part of a tensor 288
tensor 287
Symmetry principle 73
Synchronization 758
System
of equations of first approximation 693
of functions
complete 456
linearly independent 456
normalized 389, 527
of linear differential equations
adjoint 607
reducible 613
with rapidly rotating phase 761

Taylor's series 85
circle of convergence 86
radius of convergence of 86
Tensor(s) 280ff
addition of 285

affine 293f
algebra 280ff
analysis 280
antisymmetric part of 288
application to mechanics of 288ff
associated 295
Cartesian 284
component of 281
compounded from two tensors 287
contraction of 286
contravariant 292
metric 294
of rank one 293
covariant 292
metric 294
of rank one 293
covariant derivative of 305f
examples of 281ff, 288ff
field 300
in a Euclidean space 304f
in a linear space 302ff
on a manifold 308ff
general affine 291ff
in a Euclidean space 294ff
inertia 289
inner multiplication (composition) of 287
interchanging indices in 287
linear operations of 285f
lowering indices in 295
mixes 292
multiplication by a scalar of 285
numerical 292
of rank two 280, 287f
operations on 285ff
outer (tensor) miltiplication of 286
raising indices in 295
rank (order) of 284
reduction to principal axes of 288
resolution into symmetric and antisymmetric parts of 287f
skew-symmetric (antisymmetric) 287
strain 291
stress 442
symmetric part of 288
symmetric with respect to a group of indices 287
Theorem of John von Neumann 272
Theory of games 270
Time constant 720
Time-like direction 297
Topological
equivalence 43
product 676
Topology 43
Total
dissipation 711
(Gaussian) curvature 315
partial derivative 343
Trace
of a kernel 429
of a matrix 287
Transfer function 182
Transfer matrix 185
Transform 162, 195
Translation 40
of a vector 303
Transportation problem 258f
Transversal 356
Transversality condition 356, 358, 360, 361f, 363
Trefftz method 471
Triangle inequality 417
Triangular (semidiagonal)
matrix 240
system of linear equations 240
Triply-connected domain 25
Turning point 644
Two-person matrix game 270
Two-sided Laplace transformation 563
Two-way series 81

Ultraharmonic (superharmonic) oscillation 741
Umbral (dummy) index 282
Union of sets 466
Uniform deviation of functions 322
Uniform magnification 40
Unilateral constraints 363
Unit
circle 47
disk 47
function (or Heaviside step function) 183, 674
sphere 213
Unitary matrix 213
Unperturbed solution of a differential equation 684
Unstable solution of a differential equation 685
Uryson's equation 589
Utility function 257

Value of a game 272
Vandermonde determinant 507
Van der Pol's equation 747
Variation
(first) of a functional 327f
nth of a functional 369
of a function 326
one-sided 364

Variational
  method in the theory of integral equations 598f
  principle(s) 421ff, 437ff
    for conformal mapping 452ff
  problem(s) 317ff
    conditional 213, 348ff
    direct methods of solution of 454
    in parametric form 341
    isoperimetric 344ff
    Lagrange's 349
    reducible to Lagrange's problem 351ff
  reciprocal 347
  theory of eigenvalues 387ff
Vector
  differential equation 607
  line 16
  potential 26
Volterra's integral equation(s)
  nonlinear 590
  of the first and second kind 488, 579ff, 542ff
  with difference kernel 555ff
Volumetric expansion coefficient 442
Vortex
  element 64
  line 66
  point 64
Wave front 359
WBK-method (phase integral method) 643
Weak extremum 324
Weierstrass'
  function 386
  sufficient conditions for extremum of a functional 385ff
Weierstrass-Erdmann conditions 367, 414
Weight function 548
Well-conditioned
  matrix 242
  system of equations 241f
Wiener-Hopf method 564ff
Wronskian 608

Young's modulus 433, 442

Zero (or root)
  of a function 86
    multiple 86
    of order n 86
    simple 86
  of a polynomial equation 126ff
Zhukovsky's function 54

List of Symbols

Vector and Tensor Analysis:

∇ 16; Δ 21; $a_i^j{}_{.k}$ 292; g_{ij}, g^{ij} 294; Γ_{ij}^k, $\{i_k j\}$ 303, $\Gamma_{l,\,ij}$, $[ij,\,l]$, $[i_l j]$ 304; $a_{,k}^{l}$ 306

The Theory of Analytic Functions:

$\operatorname{Res}_{z=a} f(z)$ 101

Special Functions:

$\Psi_p(z)$ 136; $F(\varphi,\ k)$, $E(\varphi,\ k)$, $\Pi(\varphi,\ n,\ k)$ 141; $K(k)$, $E(k)$ 142; $\operatorname{sn}(w,\ k)$, $\operatorname{cn}(w,\ k)$, $\operatorname{dn}(w,\ k)$ 143; $K_\nu(a)$ 155; $H_\nu^{(1)}(x)$ 160; $T_n^\beta(x)$, $L_n^a(x)$, $H_n(x)$ 551; $\operatorname{ce}_n(t,\ q)$, $\operatorname{se}_n(t,\ q)$ 619; Ai (s), Bi (s) 645

Linear Algebra:

$\boldsymbol{x}$ 204; x 205; A*208; A^{-1} (0) 209; dim 210; Tr A 609

Functional Spaces. The Calculus of Variations:

$\|f\|$, $C[a,\ b]$, $C_n[a,\ b]$, $L_2[a,\ b]$ 322; $\rho(f,\ g)$ 323; δy 326; δI 327

Miscellaneous Signs: $\doteqdot$ 162; $f_1 * f_2$ 168; γ [A, (L)] 661; $\operatorname{sgn} x$ 674

TO THE READER

Mir Publishers would be grateful for your comments on the contents, translation and design of this book.
We would also be pleased to receive any suggestions you may wish to make.
Our address is:
USSR, 129820, Moscow I-110 GSP
Pervy Rizhsky Pereulok 2, Mir Publishers

Printed in the Union of Soviet Socialist Republics

OTHER BOOKS FOR YOUR LIBRARY

Descriptive Geometry
BY N. KRYLOV, P. LOBANDIYEVSKY, S. MEN

The book is written by the Soviet scientists, Candidates of Technical Sciences Nikolai Krylov and Pavel Lobandiyevsky, and engineer Şolomon Men, and is intended for students of higher technical schools majoring in building construction. A course of descriptive geometry covered by the book includes the following sections: orthogonal projections, axonometry, linear perspective and projections with numerical markings.
For better mastering of the material the last three sections include problems and solutions.
In addition to the basic course the book sets out the principles of the theory of shadows in orthogonal projections, perspective, axonometry, and this makes the manual useful also for students of mechanical engineering higher schools. The material is presented in a simple but strictly scientific form and is well illustrated. The book has seen four Russian editions and has been translated into English and Japanese.

Differential and Integral Calculus
BY N. PISKUNOV

This monograph by Prof. Nikolai Piskunov, D. Sc. (Phys. and Math.), devoted to the most important divisions of higher mathematics, deals with the following topics: number, variable function, limits, continuity of a function, derivative and differential, certain theorems on differentiable functions, investigations of the behaviour of functions, the curvature of a curve, complex numbers, polynomials, functions of several variables, applications of differential calculus to geometry in space, the indefinite integral, the definite integral, the geometric and mechanical applications of the definite integral, differential equations, multiple integrals, line and surface integrals, series, Fourier series, the equations of mathematical physics, operational calculus and certain of its applications.

Multiple Integrals, Field Theory and Series. An Advanced Course in Higher Mathematics
B. BUDAK and S. FOMIN

Covers branches of mathematics increasingly required by physicists, such as multiple, line, and improper integrals, the theory of fields, and power and trigonometric series. Based on lectures read by the authors in the physics faculty of Moscow University, the book endeavours to show the connection between the various mathematical concepts and their applications, and wherever possible their physical sense as well.

Problems in Calculus of One Variable
I. MARON

A study aid for engineering students, economists, and others, in solving problems of mathematical analysis (the functions of a variable). Suitable for home study and independent work. Most sections open with a short theoretical introduction. Answers and solutions are given. In addition to problems of the algorithmic, computational character, there are many illustrating theory and aimed at developing independent mathematical thinking. Should be used in conjunction with a suitable textbook on mathematical analysis.

OTHER BOOKS FOR YOUR LIBRARY

A. MYŠKIS, D.Sc.

LECTURES IN HIGHER MATHEMATICS

A textbook for students, revised for translation from the 3rd Russian edition.
Contents. Introduction. Quantity and Function. Analytic Geometry on a Plane. Limit. Continuity. Derivatives. Differentials. Investigating the Behaviour of Functions. Approximate Solution of Finite Functions. Interpolation. Determinants and Systems of Linear Algebraic Equations. Vectors. Complex Numbers and Functions. Functions of Several Variables. Analytic Geometry in Space. Matrices and Their Applications. Partial Derivatives. Indefinite Integrals. Definite Integrals. Differential Equations. Multiple Integrals. Series. Elements of the Probability Theory. On Modern Computers.

YA. ZELDOVICH,
MEM. USSR ACAD. SCI.

HIGHER MATHEMATICS FOR BEGINNERS

An introduction to higher mathematics for sixth formers and students of technical colleges. Can be used for home study by correspondence students. Introduces the reader to the methods of differential and integral calculus, linking them with practical problems of modern physics, radio, and electrical engineering. Contains a wealth of examples and exercises, is extensively illustrated with graphs and tables. Answers and solutions to all problems presented are provided.
Contents. Concepts of Derivatives and Integrals. Calculation of Derivatives and Integrals. Differential and Integral Calculus in the Investigation of Functions and Geometry. Functions and Graphs. Flow of Water. Radioactive Decay and Nuclear Fission. Absorption of Light. Mechanics. Thermal Motion of Molecules and the Distribution of Air Density in the Atmosphere. Electrical Circuits and Wave Phenomena in Them.